Operator Preconditioning in Hilbert Space

Tamás Kurics

PhD Thesis

Supervisor: János Karátson
Associate Professor, PhD

Mathematical Doctoral School

Director: Professor Miklós Laczkovich
Member of the Hungarian Academy of Sciences

Doctoral Program: Applied Mathematics

Director of Program: Professor György Michaletzky
Doctor of the Hungarian Academy of Sciences

Department of Applied Analysis and Computational Mathematics
Institute of Mathematics
Eötvös Loránd University, Faculty of Sciences
2010

CONTENTS

ACKNOWLEDGEMENT

I would like to express my gratitude to my esteemed supervisor Dr. János Karátson for the valuable discussions and for his inspiring lectures on functional analysis and its applications. This thesis could not have been done without his endless support and encouragement.

In the past years I have had the pleasure to meet a number of truly great people in research institutes abroad. I am indebted to Prof. Svetozar Margenov and Dr. Ivan Lirkov, Institute for Parallel Processing, Bulgarian Academy of Sciences, for their assistance and hospitality while I stayed in Sofia. I also thank Dr. Per Grove Thomsen, Department of Informatics and Mathematical Modeling, Technical University of Denmark, for the kind hospitality I received during my stay in Lyngby and for the excellent lectures he has given on stiff differential equations.

Further I would like to thank the people whom I had the pleasure to meet with at conferences or summer schools (not intended to be an exhaustive list): Dr. Maria Paz Calvo Cabrero (Universidad de Valladolid), Prof. Owe Axelsson (Uppsala Universitet), Prof. Vagn Lundsgaard Hansen (Danmarks Teknise Universitet) and Prof. Alfio Quarteroni (École Polytechnique Fédérale de Lausanne).

I would also like to thank all of my colleagues and former PhD fellows at the Department of Applied Analysis and Computational Mathematics at the Institute of Mathematics of Eötvös Loránd University for their priceless support and the great friendly atmosphere they created.

I am grateful to the support of the Deák Ferenc Scholarship provided by the Ministry of Education and Culture of Hungary.

Finally, I would like to express my deepest gratitude to my family for their support, understanding and endless patience.

OVERVIEW

The theory of elliptic partial differential equations has been a subject of extended research in the past decades. Since in general their analytic solution is not known, or difficult to handle, some kind of approximation and numerical computations are needed. The numerical solution of linear elliptic partial differential equations often involves finite element or finite difference discretization on a mesh, where the discretized system is solved by an iterative process, generally by some conjugate gradient method. The crucial point in the solution of the obtained discretized system is a reliable preconditioning, that is to keep the condition number of the systems reasonably small, possibly bounded above, no matter how the mesh parameter is chosen.

In this thesis first the investigation and numerical realization of some of the already known results of operator preconditioning are considered. The required theoretical background is summarized in the first chapters. Then we extend the scope of the theoretical results to cases that have not been covered by theory up till now. These new achievements and the numerical implementation of the considered preconditioning methods are discussed in the second part of the thesis.

In Chapter 1 we summarize the classical and modern solution methods for linear systems and turn one's attention to the importance of preconditioning. Preconditioning roughly means that one can transform the obtained linear system into a new one which is more suitable for iterative solution. This can be a purely algebraic process, but for discretized elliptic systems one can rely on the functional analytic background of the corresponding elliptic operators. This approach can be particularly advantageous, since the theory of the infinite dimensional problem in a Sobolev space is often well established, hence we can use preconditioning operators instead of preconditioning matrices. Here for the finite dimensional approximation of the original operator equation the preconditioning matrix is obtained as the projection of the corresponding operator onto the same finite dimensional subspace.

In Chapter 2 first the required background from functional analysis is summarized together with the generalized conjugate gradient methods. The choice of the preconditioner often relies on the theory of equivalent operators, which was developed in the late 1980s. The preconditioned conjugate gradient methods with equivalent preconditioners provide mesh independent linear convergence. The notion of operator equivalence

can be refined, leading to the concept of compact-equivalence, which yields superlinear mesh independent convergence. The proper treatment of the conditions in Hilbert space that ensure this favourable convergence property closes these introductory chapters.

In the second part of the thesis first we apply the theoretical background developed in the first chapters to elliptic differential operators, then we extend the theoretical results to cases that have not been covered by theory before. This part of the thesis contains the author's own contribution to the subject. We mainly deal with symmetric preconditioning, the applications of nonsymmetric preconditioners are briefly discussed in the last chapter.

In Chapter 3 we consider symmetric preconditioning for elliptic convection-diffusion equations. This is done under three different circumstances. First the case of homogeneous mixed boundary conditions is investigated, based on the papers [41, 42]. Here we compare the relation between the theoretical convergence estimate and the numerical results and we show that the convergence rate remains valid even in cases not covered by the theory. Then we extend the theory to the nonhomogeneous case by using operator pairs, relying on [40]. In contrast with finite element discretizations which fits in naturally with the Hilbert space background, there is no such abstract background for finite difference discretization, only a case-by-case study is possible. We investigate a special model problem at the end of the chapter (see [38]) and we derive a convergence estimate analogous to the finite element case.

In Chapter 4 we deal with symmetric preconditioning for elliptic systems. Here we consider decoupled symmetric preconditioners, which makes the solutions process much faster, since smaller sized independent linear algebraic systems have to be solved, hence it is easily parallelizable. First we extend the results of the previous chapter to systems (see [36]) using the already known results for equations for the case of Dirichlet boundary conditions. Then the case of mixed boundary conditions is treated using the operator pair approach with decoupled symmetric preconditioners (cf. [40]). At the end of this chapter we present a parallel algorithm (based on [39]) which was developed and implemented in cooperation with the Institute for Parallel Processing in Sofia.

Some related problems are alluded to in Chapter 5. First the application of nonsymmetric preconditioners is considered for convection-diffusion equations (cf. [37]). This is useful for problems with large convection terms, where symmetric preconditioning does not provide good enough approximation of the original elliptic operator. Then the results of the previous chapter are applied to nonlinear problems (cf. [40]). Finally a parabolic nonlinear transport system is considered. We formulate a time-dependent problem, where on each time level a nonlinear elliptic system is solved by using the preconditioning techniques developed in the preceding chapters (see [35]).

1. PRELIMINARIES

The solution of the system of linear equations

$$Ax = b, \tag{1.1}$$

where $A \in \mathbb{R}^{n\times n}$ is nonsingular and $b \in \mathbb{R}^n$, is probably one of the most studied fields in applied mathematics. Such equations naturally arise from the discretization of partial differential equations (PDE), which describe some physical phenomena governed by the laws of nature. The heat and wave propagation, electromagnetic field theory, elasto-plasticity, fluid dynamics, reaction-convection-diffusion equations, transport problems, flow models and their linearizations are the primary examples among other problems from physics, chemistry, engineering, geosciences or biology. Large sized linear systems also occur when time-dependent PDEs are discretized with respect to time with some implicit scheme. Although in this thesis this is of secondary importance, it is worth mentioning that there are a lot of other applications such as economic models or queueing systems where linear equations arise from processes not described by PDEs.

1.1 Classical solution methods for linear systems

This section is devoted to the brief description of the well-known classical solution techniques and also serves as a motivation to the further parts of this chapter. The following topics can be found in a much more detailed form in the vast literature of numerical linear algebra, we refer to the introductory textbooks [48] and [55] or the classical monographs [60] and [65].

The taxonomy of solution methods can be described very briefly. Loosely speaking, there are two types of methods: direct ones and iterative ones. A method applied to equation (1.1) is called a direct method when the exact solution – neglecting round-off errors – is available after a finite process. Another possibility is to generate a sequence of approximate solutions, which – under certain circumstances – converges to the exact solution, this is the idea of iterative methods. The borderline between these two classes is rather blurred, there exists methods that can be considered both as direct or iterative processes, gathering favourable properties from both sides, which

will be discussed later. Modifications of direct methods are also used to improve the reliability and robustness of iterative solution methods.

The main feature of the arising systems is their size, n is typically very large, systems with over a million of unknowns nowadays can be considered as routine problems. Although the size of the matrix can be huge, the number of non-zero elements is often small compared to the total number of matrix entries. This phenomenon is characteristic for systems arising from PDEs, since the discretization of the equations involves the discretization of the derivatives, which is done locally, i.e. only a certain neighbourhood of a point is used for the approximation, thus an unknown is coupled linearly with only a few number of other unknowns, making the matrix sparse. The sparsity pattern is an important property of the matrix. Sparse matrices can be stored much more efficiently than dense matrices, since most of the entries of A are zeros, and several important algorithmic procedures are implemented specifically for sparse matrices to reduce the total computational cost.

1.1.1 Direct methods

These methods are some versions of the Gaussian elimination (GE) or are matrix factorization methods that are based on that, such as the LU decomposition and its variants LDU, LUP, LDM^T. There exist slightly modified versions for symmetric, positive definite (spd) matrices such as LL^T (also known as Cholesky decomposition), LDL^T, etc. The main idea behind the GE algorithm is to replace equation (1.1) with an equivalent system (i.e. which has the same set of solution)

$$Ux = y, \tag{1.2}$$

where U is an upper triangular matrix. In one step of the GE algorithm the column entries under a diagonal element are eliminated by multiplying a row by a non-zero constant or adding such a multiplied row to another one. When the process does not break down, the result has the form (1.2). This can be solved with much less effort, since the solution of triangular systems requires n^2 flops, whilst the whole GE procedure requires a total $2n^3/3$ flops. The procedure can be used for factorizing the matrix in the form $A = LU$, if the coefficients that eliminate the corresponding elements under the diagonal are stored in a lower triangular matrix L. Then equation (1.1) can be replaced by

$$LUx = b \iff \begin{cases} Ly &= b \\ Ux &= y, \end{cases} \tag{1.3}$$

where the lower and upper triangular matrices L and U come from the elimination procedure. The arising equations can be solved by forward and backward substitution. There exist other variants of the GE algorithm developed for banded systems (arising from finite difference approximation of PDEs) and block factorization methods. Since the GE algorithm does not preserve sparsity, another important thing is to keep the level of fill-in low, which means that the unknowns can be reordered in order to preserve the sparsity pattern or at least not to lose it completely. Such processes involve graph theoretical approaches, like the reversed Cuthill–McKee algorithm or greedy coloring algorithms (see [52]), the nested dissection technique or other reorderings related to the renumbering of the nodes on the grid, like the classical red-black ordering.

Direct methods are generally robust and the required storage and process time can be predicted, which properties make them a favourable choice when reliability concerns come first (cf. [12]). Because of this, in some fields these methods are traditionally preferred. On the other hand, in two and, above all, three dimensional PDE models, very large sized system can arise and since the complexity of the GE algorithm is proportional to the cube of the number of unknowns, the application of iterative techniques simply cannot be disregarded.

1.1.2 Basic iterative methods

An iterative solution of equation (1.1) yields a sequence of approximate solutions (x_k) converging to the exact solution, often denoted by x^*. In each step, the calculation of a matrix-vector product is the costliest computation, which is generally $\mathcal{O}(n^2)$ for dense matrices, but reduced to $\mathcal{O}(n)$ for sparse matrices. For direct methods a typical complexity is $\mathcal{O}(n^2)$, such as for the banded Cholesky method, but there exist more sophisticated and efficient solvers for problems arising from the discretization of PDEs. Thus an iterative method can be competitive with direct solvers when the number of required iterations for a prescribed tolerance is less than $\mathcal{O}(n)$. The most favoured case is when the number of needed iterations is independent of the size of the problem, i.e. $\mathcal{O}(1)$. When the linear system (or family of systems indexed by the grid parameter h) arises from discretization of PDEs, then the convergence of an iterative method possessing the above property is mesh independent.

The Richardson method (also called simple iteration, fixed point iteration, etc.) is the simplest example of an iteration method. Introducing the non-zero parameter α, equation (1.1) can be equivalently transformed into

$$x = x - \alpha Ax + \alpha b,$$

leading to the linear first order stationary Richardson method

$$\begin{aligned} x_{k+1} &= (I - \alpha A)x_k + \alpha b \\ &= x_k - \alpha r_k, \end{aligned} \tag{1.4}$$

where $r_k = Ax_k - b$ is the residual vector. The iteration matrix of the method is

$$M_{\mathrm{R}}(\alpha) = I - \alpha A. \tag{1.5}$$

Assuming that A is spd and denoting its eigenvalues by $\lambda_i = \lambda_i(A)$, it is known that

$$\varrho\left(M_{\mathrm{R}}(\alpha)\right) < 1 \iff 0 < \alpha < \frac{2}{\lambda_{\max}}. \tag{1.6}$$

The value of the optimal parameter – which minimizes the spectral radius of $M_{\mathrm{R}}(\alpha)$ – is given by the formula

$$\alpha_{\mathrm{opt}} = \frac{2}{\lambda_{\max} + \lambda_{\min}}. \tag{1.7}$$

Then the convergence factor of the Richardson method (1.4) using the optimal parameter given in (1.7) is

$$\varrho\left(M_{\mathrm{R}}(\alpha_{\mathrm{opt}})\right) = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} = \frac{\kappa - 1}{\kappa + 1}, \tag{1.8}$$

where κ denotes the spectral condition number of A. Iteration (1.4) can also be obtained from a special splitting of A:

$$\begin{aligned} Ax = b &\iff \big(\omega I - (\omega I - A)\big)x = b \iff \omega x = (\omega I - A)x + b \\ &\iff x = x - \alpha Ax + \alpha b, \quad \text{where } \alpha = \omega^{-1}. \end{aligned}$$

The general idea is the following. Let us split the matrix A into two parts

$$A = P - N, \tag{1.9}$$

where P is invertible. Then equation (1.1) can be rewritten as

$$x = P^{-1}Nx + P^{-1}b, \tag{1.10}$$

or alternatively

$$x = x - P^{-1}(Ax - b). \tag{1.11}$$

This gives rise to the iteration procedure of the form

$$x_{k+1} = Mx_k + v, \tag{1.12}$$

where $M = P^{-1}N$ and $v = P^{-1}b$, or alternatively

$$x_{k+1} = x_k - P^{-1}r_k. \tag{1.13}$$

Here and hereafter vector coordinates and algorithms calculating such coordinates are not considered, thus the subscript denotes simply the numbering of the elements in the sequence of vectors. The following well-known result (see e.g. in [55]) gives a necessary and sufficient condition for the convergence of iteration (1.12).

Proposition 1.1. *If $A = P - N$, $M = P^{-1}N$ and $v = P^{-1}b$, then the sequence (x_k) generated by the iteration* (1.12) *converges for all initial vectors x_0 if and only if $\varrho(M) < 1$, where $\varrho(M)$ is the spectral radius of M.*

The classical linear iterative methods are the Jacobi and Gauss–Seidel iterations and their relaxed versions. Let us consider the decomposition $A = L + D + U$, where D is a diagonal matrix consisting of the diagonal of A, further L and U are the lower and upper triangular parts of A (excluding the diagonal itself), respectively.

Let us assume that there are no zero entries in the diagonal of A. If $P = D$ is chosen in the splitting (1.9), then

$$x_{k+1} = -D^{-1}(L + U)x_k + D^{-1}b \tag{1.14}$$

is called the Jacobi method, and when $P = L + D$ is chosen then

$$x_{k+1} = -(L + D)^{-1}Ux_k + (L + D)^{-1}b \tag{1.15}$$

is called the Gauss–Seidel method. The iteration matrices are

$$M_{\mathrm{J}} = -D^{-1}(L + U) = I - D^{-1}A, \tag{1.16}$$

$$M_{\mathrm{GS}} = -(L + D)^{-1}U = I - (L + D)^{-1}A. \tag{1.17}$$

The inversion in these matrices does not need to be executed, since the iterations can be written in a convenient coordinatewise form. In the relaxation methods a relaxation (or damping) parameter ω is involved. The corresponding schemes are called JOR (over-

relaxation) and SOR (successive over-relaxation) methods with iteration matrices

$$M_{\mathrm{J}}(\omega) = \omega M_{\mathrm{J}} + (1-\omega)I = I - \omega D^{-1}A, \tag{1.18}$$

$$M_{\mathrm{GS}}(\omega) = (D+\omega L)^{-1}\left((1-\omega)D - \omega U\right). \tag{1.19}$$

Other methods are also common, such as the symmetrized form of the relaxed Gauss–Seidel, which is called SSOR (symmetric successive over-relaxation) method, the alternating direction implicit method (also known as Peaceman–Rachford method) or the cyclic reduction method. Block iterations could be also considered. The boom of matrix theory in the mid 20th century is closely connected with the profound study of these methods. Several convergence results were obtained by introducing special classes of matrices, those arising in practice, such as M-matrices (introduced by Ostrowski), Stieltjes matrices, nonnegative and irreducible matrices, together with splittings of A with special properties in (1.9) such as the regular splitting. It has been shown that the classical iterations (1.14) and (1.15) are convergent for strictly diagonally dominant matrices and for M-matrices as well. As for the SOR method, the relaxation parameter has to satisfy the inequality $0 < \omega < 2$, but for spd matrices this condition is also sufficient for convergence, due to the theorems of Kahan and Ostrowski. When M_{J} in (1.16) happens to be nonnegative, then methods (1.14) and (1.15) are equi-convergent, that is they either both converge or both diverge, a result known as the Stein–Rosenberg theorem. Regarding the speed of convergence, it has been shown, for instance, that for the class of strictly diagonally dominant matrices the Gauss–Seidel iteration is at least as good as the Jacobi iteration, and in the case of block-tridiagonal matrices the Gauss–Seidel iteration performs considerably faster than the Jacobi method. The optimal choice of the acceleration parameter has been also investigated in several circumstances, often involving demanding eigenvalue analysis. For the precise formulation of the theorems and proofs we only refer to the books [7, 52, 60, 65], where some interesting historical remarks can also be found. For further reading for the classical matrix classes that are related to discretized PDEs, we refer to [29, 30, 65].

1.2 Modern iterative methods

The classical iteration schemes considered in Subsection 1.1.2 were stationary linear iterative processes of first order, which means that for the computation of the new approximate solution x_{k+1} only the previous vector x_k was used. Furthermore, the same iterative process was used to calculate the next vector, the same scheme was applied repeatedly, using the same parameter (if there were any), since a fixed point iteration is in the background. The main drawback of these methods is that generally

it is very difficult to estimate the convergence factor without *a priori* information and for many practical problems the convergence of these methods is very slow.

A possible remedy is to allow nonstationary methods, where the parameters involved are chosen dynamically, satisfying some optimality properties, generally of a geometric nature, i.e. minimizing the error in each step in some subspace or satisfying some kind of orthogonality conditions. For the construction of such subspace of constraints one may use all or some of the previous approximate vectors. These optimality requirements can also be satisfied in infinite dimensional inner product spaces, thus some of these methods can be generalized for solving operator equations in Hilbert space. Another possibility is to improve the spectral bounds of A, that is to make it better conditioned, an idea that has become one of the most crucial step in the solution process. These methods were investigated first in the early 1950s, but soon fell into oblivion, because their first implementations were not competitive with the then widely used overrelaxation methods. However, they were paid attention again in the early 1970s, as the revolutionary growth of computer-aided numerical computations made it easier to implement and run those algorithms efficiently. The related methods, the so-called Krylov subspace iterations have become standard topics in numerical textbooks.

1.2.1 Preconditioning

Let us revisit the Richardson method (1.4) and comment the convergence results (1.7)-(1.8). The calculation of the optimal α requires exact information about the extremal eigenvalues of A. They are not known generally, but usually some estimation is available, thus it is possible to choose α with property (1.6). But for ill-conditioned systems, when the interval containing the eigenvalues is large, the convergence factor – even with the optimal parameter – is close to 1, which provides very slow convergence. This situation is typical for systems arising from the discretization of elliptic boundary value problems (BVP). If the elliptic PDE is of order $2m$, then the condition number behaves like $\mathcal{O}(h^{-2m})$, where h is the mesh parameter. For second order elliptic PDEs this means that

$$\kappa_h \sim \mathcal{O}(h^{-2}) \to \infty, \text{ when } h \to 0, \tag{1.20}$$

regardless of the dimension of the domain. The smaller the discretization parameter h is, the higher the required number of iteration steps is needed, which is a major drawback, considering that the larger size of the problem itself implies the increase of computational costs. The remedy is the following: modify algorithm (1.4) and apply

the iteration scheme to a new equation

$$P^{-1}Ax = P^{-1}b, \tag{1.21}$$

where P is some invertible matrix. The idea comes from the observation that the eigenvalue distribution of $P^{-1}A$ may be more favourable than of A, thus the iteration could be considerably accelerated. This idea, when one tries to squeeze the spectrum of A into a small region of the complex field in order to reduce the condition number of the system (1.1) is called preconditioning, and the matrix P is called preconditioner. In this case, the modified iteration scheme for the Richardson method is given by

$$\begin{aligned} x_{k+1} &= (I - \alpha P^{-1}A)x_k + \alpha P^{-1}b \\ &= x_k - \alpha P^{-1}(Ax_k - b) \\ &= x_k - \alpha P^{-1}r_k \end{aligned} \tag{1.22}$$

with the iteration matrix

$$M_{\mathrm{R},P}(\alpha) = I - \alpha P^{-1}A. \tag{1.23}$$

The convergence factor (1.8) shows that the more the eigenvalues of $\alpha P^{-1}A$ are clustered around 1, the more efficient the preconditioning is.

The Jacobi and the Gauss–Seidel methods – and the classical relaxation methods – can also be considered as stationary preconditioned Richardson iterations: for instance, the choices $\alpha = 1$, $P = D$, and $\alpha = 1$, $P = L + D$ give back the iteration matrices (1.16) and (1.17), respectively.

The matrix $P^{-1}A$ of the preconditioned system is never formed explicitly (unless P^{-1} is known exactly), since the inversion of P and matrix-matrix products would be too expensive. Instead of this, in each step of algorithm (1.22) an auxiliary equation has to be solved:

Algorithm 1.2 (Preconditioned Richardson method)**.**

1.) Let $x_0 \in \mathbb{R}^n$ be arbitrary, $r_0 = Ax_0 - b$;

2.) For given x_k

2a.) solve $Py_k = r_k$;

2b.) $x_{k+1} = x_k - \alpha y_k$;

2c.) $r_{k+1} = r_k - \alpha A y_k$.

Preconditioning is thus nothing else than transforming the system (1.1) equivalently into the system (1.21) which has more favourable properties for iterative so-

lution. Equation (1.21) is also called left-preconditioning, but system (1.1) could be preconditioned from the right as

$$AP^{-1}y = b, \quad x = P^{-1}y, \tag{1.24}$$

or simultaneously from both sides

$$P_1^{-1}AP_2^{-1}y = P_1^{-1}b, \quad x = P_2^{-1}y, \tag{1.25}$$

which is called split or centered preconditioning. When a preconditioner is chosen, one has to keep the following natural criteria in view:

1. $P^{-1}A$ should be considerably better conditioned than A;
2. (a) the preconditioned system should be easy to solve, that is the solution of systems with P should not be costly;
   (b) the construction of the preconditioner P should be easy and cheap;

An additional requirement can be the following:

   (c) P should be close to optimal in the sense that the number of required iterations to reach a prescribed tolerance level should be independent of the size of the system.

Note that these criteria are conflicting, the optimal choice $P = A$ obviously satisfies the first criterion, but not the second one, whilst the choice $P = I$ makes the solution of the auxiliary systems trivial but does not make the convergence faster. The proper choice of the preconditioner is thus not obvious at all, it can strongly depend on the structure of A or on the PDE itself hiding behind the discretized system. Although there is no universal way of obtaining good preconditioners for every problem, generally a preconditioner is not far from being good if the spectrum of the obtained system is small enough and the preconditioned matrix is close to a normal matrix (cf. [12]).

There are two approaches to choose preconditioners. The first one disregards the original problem from which the linear system is originated. This can happen for several reasons, usually when complete information about the problem is not available or it would be difficult to use. In this case some universal preconditioner is needed, which can be far from being optimal, but can be applied to a wider class of problems. This approach can use only the information that can be obtained from the matrix itself and the preconditioner is constructed via an algebraic process, thus it is called algebraic preconditioning.

In the second approach the goal is to choose an optimal or close to optimal preconditioner for a special class of problems. This approach is used when the continuous model described by a PDE lying behind the linear system is well understood. All the available information that can be gathered from the model properties can be used to obtain a good preconditioner, which is usually derived from the discretization of another, simpler PDE that is close to the original continuous problem in some sense. This problem-specific way of obtaining preconditioners is called continuous or functional preconditioning to emphasize the continuous model behind. Here the preconditioning process can take place on the operator level as well, where the corresponding operators act between Sobolev spaces, involving elements of functional analysis, usually Hilbert space theory. In this case the preconditioning matrix is considered as the projection of the preconditioning operator onto the same finite dimensional discretization space, where the original operator was discretized. For this reason this technique is also referred as Sobolev or operator preconditioning. This approach and some results from operator preconditioning is the main topic of this thesis.

For completeness the most common algebraic preconditioning techniques are summarized here to close this subsection.

Incomplete factorization methods. As explained in Subsection 1.1.1, the LU factorization of A may be unsatisfactory due to the high number of fill-ins, destroying the favourable sparsity pattern of A and increasing the computational cost. The idea behind the incomplete factorization methods is to preserve (some of) the sparsity, i.e. the preconditioner P is chosen to be the approximate decomposition of A:

$$P = \widetilde{L}\widetilde{U}, \tag{1.26}$$

where $\widetilde{L}$ and $\widetilde{U}$ are lower and upper triangular matrices approximating L and U, respectively, where $A = LU$. When no fill-in is allowed, that is only those elements are calculated during the GE process where the original entry differs from zero, then the process is called the ILU(0) method. Similarly, for symmetric matrices the corresponding incomplete Cholesky decomposition is called IC or IC(0). To improve the accuracy, some fill-in can be accepted. In this case, to every matrix element a fill-in level is assigned, which is being modified during the algorithm. When this level exceeds some fixed number $p \in \mathbb{N}$, then the corresponding element is set to be zero. The resulting ILU(p) algorithm combined with some reordering process is a very efficient way of obtaining preconditioners, even for small values of p. Another way to improve the quality of the factorization is to enforce the preconditioner to have the same row sums as the original matrix

by adding the dropped fill-ins to the diagonal elements. This is the idea of the modified incomplete factorization methods (MILU, MIC).

Sparse approximate inverses. Here a sparse matrix P is computed as the direct approximation of A^{-1}, in other words $AP \approx I$, where some *a priori* sparsity pattern or bandwidth is given. In one of the main approaches this approximate inverse can be obtained by solving the Frobenius-norm minimization problem

$$\min_{P \in \mathcal{S}} \|I - AP\|_F$$

leading to least-square problems, where $\mathcal{S}$ is a class of matrices possessing some given sparsity pattern.

Multigrid and algebraic multilevel methods. The multigrid method (MG) is an iterative solution method constructed for systems arising from either finite difference (FDM) or finite element (FEM) discretization of elliptic PDEs with optimal computational complexity $\mathcal{O}(n)$. The idea behind MG is that the classical iteration schemes could damp the error associated with high frequency components, although their overall performance in damping the total error is weak. Given an initial guess x_0, a few number of iteration steps of the relaxed Jacobi, SOR or SSOR methods can smooth out those components from the error significantly. If $r = Ax_0 - b$ is the residual vector, then $x_1 = x_0 - v$ already satisfies $Ax_1 = b$, where v is the solution of $Av = r$. The second system can be solved on a coarser grid, and (keeping in mind that the high frequency components of the higher dimensional vector x_0 are already obtained), the solution then could be interpolated onto the original grid and the initial guess can be corrected, decreasing the error in the low frequency components as well. This step is usually referred as coarse grid correction. The process can be extended for more than two grids, calling the algorithm recursively and solving the linear equation exactly on the coarsest grid only.

The MG method then consists of a finite sequence of grids $\mathcal{T}_1 \subset \mathcal{T}_2 \subset \ldots \subset \mathcal{T}_m$, linear operators $P_{\ell-1}^{\ell} : \mathcal{T}_{\ell-1} \to \mathcal{T}_\ell$ and $R_\ell^{\ell-1} : \mathcal{T}_\ell \to \mathcal{T}_{\ell-1}$ called prolongation and restriction operators and smoothers $S_\ell : \mathcal{T}_\ell \to \mathcal{T}_\ell$ ($\ell = 2, \ldots, m$). In the FEM case, the grids are nested triangulations of the domain (explaining the notation $\mathcal{T}_\ell$) and other relations hold between the prolongation and restriction.

Algorithm 1.3 (MGC(ℓ, x, b))**.**

1.) If $\ell = 1$, then $x = A_1^{-1}b$; % solution on the coarsest grid

2.) else

2a.) $x = S_\ell^{\nu_1} x$; % pre-smoothing ν_1 times

2b.) $r = R_\ell^{\ell-1}(A_\ell x - b)$; % restriction of the residual

2c.) $v = 0$; % starting guess for the correction term

2d.) for $i = 1 : \gamma$ % recursive call γ times

2di.) MGC$(\ell - 1, v, r)$; % calculation of the correction term

2e.) $x = x - P_{\ell-1}^{\ell} v$; % coarse grid correction

2f.) $x = S_\ell^{\nu_2} x$. % post-smoothing ν_2 times

Then the MG method can be obtained by calling this routine on the finest grid with initial guess $x_0 = 0$. The parameter γ shows how many times each level is visited, the particularly interesting cases are $\gamma = 1$, called V-cycle and $\gamma = 2$, called W-cycle. This procedure is also called geometric multigrid method (GMG), because of the physical presence of the grids.

Algebraic multilevel methods can be considered as the generalization of the MG method. Here the grids are obtained from the graph of A and the refinement is made as selecting subsets of unknowns, without any geometry in mind. This method – called algebraic multigrid (AMG) – can be extended when the preconditioners are based on the recursive block-partitioning of the matrix associated with some hierarchical partitioning of matrix graph; this is the starting point of the algebraic multilevel iterations (AMLI). This is more general than the AMG method, because of the lack of multigrid components smoothing, restriction and prolongation, and is based on some approximation of the Schur complement.

Domain decomposition methods. This method is developed for the solution of PDEs discretized on a complicated domain. The original computational domain Ω is decomposed into subdomains Ω_i $(i = 1, \ldots, m)$, which may or may not overlap. Then the original problem can be reformulated on each subdomain, resulting a family of smaller sized problems coupled through the values of unknowns lying on the common boundaries or overlapping parts of the subdomains. Treating them via an iterative process, the coupling can be relaxed and in each iteration step the smaller problems can be solved independently, thus this method is easily parallelizable on a multiprocessor architecture. Methods where this idea appeared were initially introduced by Schwarz in the 19th century, thus they are also known as Schwarz preconditioners, the method is referred to as Schwarz alternating procedure.

These algebraic methods are discussed in much more details in the books [7] and [52], among other techniques not mentioned here. The idea of the MG method goes back to Fedorenko's early papers, for a detailed description of the topic see [25, 26]. Incomplete factorization methods and the sparse approximate inverse technique are discussed in the comprehensive survey [12]. The Schwarz method is also treated in [49].

1.2.2 Krylov subspace methods

The Richardson method can be accelerated further if the parameter α in (1.4) is chosen dynamically. This can be done by either minimizing some functional related to the equation or satisfying appropriate orthogonality properties. These approaches lead to essentially the same family of iterative methods, giving the opportunity for further generalizations.

Assume that A is spd and let us define the quadratic functional $\Phi : \mathbb{R}^n \to \mathbb{R}$ as

$$\Phi(x) = \frac{1}{2} \langle Ax, x \rangle - \langle b, x \rangle , \tag{1.27}$$

where $\langle \cdot, \cdot \rangle$ denotes the standard inner product on $\mathbb{R}^n$, generating the eucledian norm $\|x\| = \sqrt{\langle x, x \rangle}$. This is called the energy functional of equation (1.1). Denoting the solution of (1.1) by x^*, a simple calculation shows that the quadratic functional Φ has a unique minimum attained in x^*. Thus finding the solution of (1.1) is equivalent to finding the minimizer of the functional Φ.

If only the last approximation x_k is used, then the generic step of the algorithm has the form

$$x_{k+1} = x_k + \alpha_k d_k,$$

where the new approximation is a correction of x_k in the search direction d_k. When the vector d_k points towards the minimal slope of Φ, that is

$$\partial_{d_k} \Phi(x_k) = \min\{\partial_d \Phi(x_k) : \ d \in \mathbb{R}^n, \ \|d\| = 1\},$$

then a nonstationary algorithm can be obtained which is called gradient or steepest descent method. Calculating the directional derivative of Φ, it is easy to see that $\partial_d \Phi(x) = \langle Ax - b, d \rangle$, which is minimal if $d = -(Ax - b) = -r$. Thus the iteration has the form

$$x_{k+1} = x_k - \alpha_k r_k.$$

This is the Richardson iteration (1.4) again. The case $\alpha_n \equiv \alpha$ has been already discussed there, but now it is possible to chose the parameter α_n to be optimal in the

nth step:

$$\Phi(x_n - \alpha_n r_n) = \min_{\alpha>0} \Phi(x_n - \alpha r_n).$$

Since the latter is a minimization problem of a quadratic polynomial, it is easy calculate that the optimal parameter is

$$\alpha_n = \frac{\|r_n\|^2}{\langle Ar_n, r_n\rangle}. \tag{1.28}$$

Note that the dynamically chosen optimal parameter does not require any estimate of the extremal eigenvalues.

Algorithm 1.4 (Gradient method)**.**

1.) Let $x_0 \in \mathbb{R}^n$ be arbitrary, $r_0 = Ax_0 - b$;

2.) For given x_k

2a.) $r_k = Ax_k - b$.

2b.) $\alpha_k = \dfrac{\|r_k\|^2}{\langle Ar_k, r_k\rangle}$;

2c.) $x_{k+1} = x_k - \alpha_k r_k$;

Another approach is to satisfy some orthogonality constraints. This leads to the very general framework of projection methods, described in [52]. Following the notations used there, here in each step the new approximate solution $\widetilde{x}$ is located in an affine subspace $x + K$ in such a way that the residual vector r is orthogonal to another subspace L having the same dimension as K. A projection method is said to be orthogonal if $L = K$, but in other cases the subspace of constraints L can be completely unrelated to the search subspace K. In the case of orthogonal projection methods the orthogonality conditions are called Ritz–Galerkin conditions. When L is different from K, then it is an oblique projection method with orthogonality constraints referred as Petrov–Galerkin conditions. If those subspaces are one dimensional in each step, say $K = \operatorname{span}\{v\}$ and $L = \operatorname{span}\{w\}$, then for a given vector x, the new approximate solution has the form $\widetilde{x} = x - \alpha v$, satisfying the orthogonality condition $\langle A\widetilde{x} - b, w\rangle = 0$. From these conditions the value of α can be easily calculated:

$$\alpha = \frac{\langle r, w\rangle}{\langle Av, w\rangle}.$$

If in the kth step v and w are set to be r_k, then this gives back the optimal value (1.28) in the gradient method. Therefore the gradient method is an orthogonal projection method, where the subspaces K and L are the one dimensional subspaces spanned by

the residual vector. There are other popular choices like $v := r_k$ and $w := Ar_k$ for nonsymmetric positive definite matrices, which is called the minimal residual iteration.

A specific choice of the sequence of subspaces leads to the methods of conjugate gradients, a family of algorithms that has been selected into the top ten algorithms of the century. Conjugate means that the descent directions are chosen to be mutually A-orthogonal, i.e. the new approximation is searched in the direction of d_k, where $\langle Ad_i, d_j\rangle = 0$ $(i \neq j)$. In the standard methods the search subspace in the kth step is

$$\mathcal{K}_k \equiv \mathcal{K}_k(r_0) := \operatorname{span}\{r_0, Ar_0, \ldots, A^k r_0\},$$

the so-called Krylov subspace, generated by the initial residual r_0. In terms of projection methods, the search subspace in the kth step is $\mathcal{K}_k$, the subspace of constraints is either $\mathcal{K}_k$ or $A\mathcal{K}_k$. The process can also be considered as the minimization of the functional Φ over the affine subspace $x_0+\mathcal{K}_k$, or the minimization of the residual (leading to methods like the GMRES). Applying the standard method for spd matrices the exact solution can be obtained – in the absence of round-off errors – in at most n steps, thus it can be considered as a direct method.

The algorithm was introduced by Hestenes and Stiefel in [28], and was considered first as a direct method, but later it has been discovered that the algorithm provides good approximation with far fewer iteration steps. Its three term recurrence form for spd matrices was invented by Lanczos (see [43, 44]). In the past 40 years a number of related methods have been discovered and investigated, such as generalized conjugate gradient methods or their variants for nonsymmetric or indefinite matrices.

The standard conjugate gradient method for spd matrices is as follows:

Algorithm 1.5 (Conjugate gradient method (CG))**.**

1.) Let $x_0 \in \mathbb{R}^n$ be arbitrary, $d_0 = r_0 = Ax_0 - b$;

2.) For given x_k, d_k and $r_k = Ax_k - b$, we let

2a.) $x_{k+1} = x_k + \alpha_k d_k$, where $\alpha_k = -\dfrac{\langle r_k, d_k\rangle}{\langle Ad_k, d_k\rangle}$,

2b.) $d_{k+1} = r_{k+1} + \beta_k d_k$, where $\beta_k = \dfrac{\|r_{k+1}\|^2}{\|r_k\|^2}$;

Note that the method is parameter-free, and for indefinite matrices $\langle Ad_k, d_k\rangle$ may be zero, even if $d_k \neq 0$, so the standard CG algorithm can break down. The construction of the CG algorithm yields the following optimality property (cf. [7, Chap. 13]).

Proposition 1.6. *Let $e_k = x_k - x^*$ be the error vector and $\mathbb{P}_k^1 = \{p_k \in \mathbb{R}[x] : \deg p_k \leq k,\ p_n(0) = 1\}$. Then*

$$\|e_k\|_A = \min_{p_k \in \mathbb{P}_k^1} \|p_k(A)e_0\|_A \,, \tag{1.29}$$

where $\|e_k\|_A = \sqrt{\langle Ae_k, e_k\rangle}$.

Remark 1.7. If the eigenvectors of A form an orthonormal basis (which does hold for symmetric matrices), the bound from the optimality property (1.29) can be further estimated as

$$\frac{\|e_k\|_A}{\|e_0\|_A} \leq \min_{p_k \in \mathbb{P}_k^1} \max_{\lambda \in \sigma(A)} |p_k(\lambda)| \,.$$

The spectrum of an spd matrix is real and bounded by its extremal eigenvalues. The upper bound above can be estimated by using Chebyshev polynomials of first kind and we get the following linear convergence theorem.

Theorem 1.8. *If A is spd, then the standard CG algorithm 1.5 yields*

$$\left(\frac{\|e_k\|_A}{\|e_0\|_A}\right)^{1/k} \leq 2^{1/k}\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \qquad (k = 1, \ldots, n), \tag{1.30}$$

where $\kappa = \kappa(A)$ is the spectral condition number of A.

One of the most important properties of the conjugate gradient method is superlinear convergence, first proved in [27], where the CGM was formulated in Hilbert space. The result has been partially extended for nonsymmetric systems which are diagonalizable and have positive symmetric part (i.e. $A + A^* > 0$), see in [4, 6, 7]. Early results on the CGM in Hilbert space can be found in [17, 27, 63], other Hilbert space methods are also summarized in [47].

Consider the matrix A in (1.1) as the perturbation of the identity matrix, that is

$$A = I + C,$$

and denote by $\lambda_k = \lambda_k(C)$ $(k = 1, \ldots, n)$ the ordered eigenvalues of C, that is $|\lambda_1(C)| \geq \ldots \geq |\lambda_n(C)|$. Then the CG method yields

$$\left(\frac{\|e_k\|_A}{\|e_0\|_A}\right)^{1/k} \leq \frac{2}{k}\sum_{i=1}^{k}\left|\frac{\lambda_i(C)}{1+\lambda_i(C)}\right| \qquad (k = 1, \ldots, n), \tag{1.31}$$

where in the bound only separate eigenvalues are involved. If $|\lambda_i| < 1/3$, then for sufficiently large k the convergence factor is smaller than 1 and decreases, see for more details in [7, Chap. 13.] and [9]. For spd matrices the following result holds.

Theorem 1.9. *(cf. [11]) The standard CG algorithm 1.5 yields*

$$\left(\frac{\|e_k\|_A}{\|e_0\|_A}\right)^{1/k} \leq 2\left\|A^{-1}\right\|\left(\frac{1}{k}\sum_{i=1}^{k}|\lambda_i(C)|\right) \qquad (k=1,\ldots,n), \tag{1.32}$$

where $|\lambda_1(C)| \geq |\lambda_2(C)| \geq \ldots \geq |\lambda_n(C)|$ *are the ordered eigenvalues of* C*.*

When the eigenvalues of C accumulate at the origin, the upper bound in (1.32) decreases as k increases, resulting superlinear convergence.

A lot of other Krylov subspace methods exist, such as the Arnoldi method (introduced for transforming a matrix into Hessenberg form for eigenvalue estimation), the Arnoldi method for linear systems (called FOM), the Lanczos method (the simplified version of Arnoldi's method for symmetric matrices), and the ones based on the residual minimization approach. For nonsymmetric systems GMRES – a generalized residual minimizing algorithm which does not break down even for indefinite matrices unless it has already converged – is widely used, first introduced in [53]. Further methods are MINRES, CGR (introduced in [19]), Orthomin, Orthodir, or the further generalized Bi-CG, Bi-CGSTAB (cf. [58]). There exists hybrid methods, and truncated or restarted versions of these algorithms can also be considered. Generalized CG algorithms that are suitable for nonsymmetric matrices were introduced in [6] (GCG-LS) and in [14, 62] (called CGW method), among several others.

For nonsymmetric systems the equation can also be symmetrized by considering the normal equation

$$A^T A x = A^T b$$

and a method for symmetric problems can be applied, the standard CG algorithm for instance (CGN method), although the amount of work in each iteration step doubles and the rate of convergence slows down considerably.

The algorithms of the generalized conjugate gradient methods that will be used in further chapters and the related convergence theorems are listed in Section 2.2. The interested reader may find much more details about these methods, their preconditioned versions and convergence theorems in the monographs [7, Chaps. 11-13], [52, Chaps. 5-9] and [59]. A short summary can be found in [4, 48, 49].

2. SOME BACKGROUND ON OPERATOR PRECONDITIONING

In this chapter the theoretical background of operator preconditioning is summarized. For a given PDE one approximates the differential operator by a simpler (e.g. symmetric) differential operator to obtain an efficient preconditioner on the operator level. Then the discretization of the preconditioning operator is used as a preconditioning matrix for the corresponding discretized system, which is solved by some conjugate gradient method. These methods are discussed in Section 2.2. One of the main features of these algorithms is superlinear convergence which is – under certain circumstances (see Sections 2.3-2.4) – mesh independent, i.e. independent of the chosen FEM subspace and the size of the discretized system. Namely, the convergence factor can be estimated by some characteristic feature of the preconditioning operator.

2.1 Basic notions

In this section some useful concepts are summarized from functional analysis and from the theory of Sobolev spaces, which will be used throughout in further chapters. The Hilbert space setting is also suitable to list the generalized conjugate gradient methods and the related convergence theorems that will be used later on.

2.1.1 Prerequisites from functional analysis

Let H_1, H_2 be Hilbert spaces, then the space of bounded linear operators mapping H_1 into H_2 is denoted by $B(H_1, H_2)$. For $H_1 = H_2$, let $B(H_1) := B(H_1, H_1)$. The topological dual space of H – that is, the space of bounded linear functionals – is denoted by H^*.

Theorem 2.1 (Riesz' representation theorem)**.** *Let H be a Hilbert space, $\varphi \in H^*$ be a bounded linear functional. Then there exists a uniquely determined $y \in H$ such that*

$$\varphi(x) = \langle x, y \rangle \qquad \forall\, x \in H,$$

moreover, $\|\varphi\| = \|y\|$.

Theorem 2.2 (Lax–Milgram lemma)**.** *(cf. [5]) Assume that H is a Hilbert space, $a : H \times H \to \mathbb{C}$ is a bounded and coercive sesquilinear functional and $\varphi \in H^*$. Then there exists a uniquely determined $y \in H$ such that*

$$a(x, y) = \varphi(x) \qquad \forall\ x \in H.$$

Definition 2.3. If $A \in B(H)$ is a bounded linear operator, then there exists a uniquely determined operator $A^* \in B(H)$, called the adjoint of A, such that

$$\langle Ax, y\rangle = \langle x, A^*y\rangle \qquad \forall\ x, y \in H.$$

An operator A is self-adjoint if $A = A^*$, and normal if it commutes with its adjoint, i.e. $AA^* = A^*A$.

Definition 2.4. Let H_1 and H_2 be Hilbert spaces. A linear operator $K : H_1 \to H_2$ is compact if it maps bounded sets into relative compact sets.

Remark 2.5. An operator K is compact if and only if for every bounded sequence $(x_n) \subset H_1$ a convergent subsequence can be selected from $(Kx_n) \subset H_2$.

A compact linear operator is bounded and the set of compact operators is a subspace in $B(H_1, H_2)$. Moreover, if $H = H_1 = H_2$, the vector space of compact operators form a two-sided ideal in $B(H)$.

Theorem 2.6 (Hilbert–Schmidt)**.** *(cf. [15, 66]) Let H be an infinite dimensional complex separable Hilbert space, $A \in B(H)$ be a compact normal operator. Then*

1. *the spectrum $\sigma(A) \subset \mathbb{C}$ of A is a countable set and $\sigma(A) = \bigcup_{k\in\mathbb{N}} \{\lambda_k(A)\} \cup \{0\}$, where $\lambda_k(A)$ are the eigenvalues of A;*
2. *the set of eigenvalues has the zero as the only limit point;*
3. *for any non-zero eigenvalue of A the corresponding eigenspace is finite dimensional;*
4. *the eigenvectors can be chosen to form a complete orthonormal basis in H.*

Many of the operators that occur in applications (in the theory of PDEs or in mathematical physics) are not bounded. Here some basic definitions and theorems are summarized that will be used later on.

Definition 2.7. Let $A : D(A) \subset H \to H$ be a densely defined linear operator on H. Let $D(A^*) = \{y \in H :\ \exists\ y^* \in H$ such that $\langle Ax, y\rangle = \langle x, y^*\rangle\ \forall\ x \in D(A)\}$. Then for each $y \in D(A^*)$, we define $A^*y := y^*$. The linear operator A^* is well-defined and called the adjoint of A.

Definition 2.8. An operator $A : D(A) \subset H \to H$ is symmetric if A is densely defined and

$$\langle Ax, y\rangle = \langle x, Ay\rangle \qquad \forall\, x, y \in D(A).$$

Remark 2.9. Let A be a densely defined operator. Then the following statements are equivalent:

1. A is symmetric;
2. $A \subset A^*$, which means that $D(A) \subset D(A^*)$ and $A^*\big|_{D(A)} = A$, that is A^* is an extension of A;
3. $\langle Ax, x\rangle \in \mathbb{R}$ for any $x \in D(A)$.

Definition 2.10. A densely defined operator A is self-adjoint if $A = A^*$.

According to Remark 2.9, a densely defined operator is self-adjoint if $A^* \subset A$.

Definition 2.11. An operator $A : D(A) \subset H \to H$ is called strongly positive if there exists some positive constant $m > 0$, such that

$$\langle Ax, x\rangle \geq m \|x\|^2 \qquad \forall\, x \in D(A). \tag{2.1}$$

Proposition 2.12. *(cf. [15]) Let A be a symmetric operator on H and assume that A is surjective, i.e. $R(A) = H$. Then A is self-adjoint.*

The following proposition, which states the converse in some sense, is a consequence of the closed range theorem, see [64, Chap. VII].

Proposition 2.13. *Let H be a Hilbert space, A be a densely defined closed linear operator and assume that*

$$\operatorname{Re}\langle Ax, x\rangle \geq m \|x\|^2 \qquad \forall\, x \in D(A)$$

for some positive constant $m > 0$. Then A^* is surjective.

An operator is closed if its graph $G(A) = \{(x, Ax) :\ x \in D(A)\}$ is a closed subset of $H \times H$. The adjoint of a densely defined operator is closed, thus with the combination of Proposition 2.12 and 2.13, the following – frequently used – consequence can be obtained.

Corollary 2.14. *Let A be a symmetric operator satisfying* (2.1)*. Then A is self-adjoint if and only if $R(A) = H$.*

Definition 2.15. Let $A : D(A) \subset H \to H$ be a symmetric operator which is strongly positive. Then

$$\langle x, y \rangle_A := \langle Ax, y \rangle \qquad \forall\, x, y \in D(A)$$

defines an inner product on $D(A)$. Let $H_A := [D(A), \langle \cdot, \cdot \rangle_A]$, i.e. the completion of $D(A)$ under the inner product $\langle \cdot, \cdot \rangle_A$. The Hilbert space H_A is called the energy space of A endowed with the energy inner product $\langle \cdot, \cdot \rangle_A$.

Proposition 2.16. *(cf. [66]) There exists a continuous linear map from H_A to H, which is injective, i.e. the energy space H_A can be identified with a subspace of H.*

This identification justifies the set inclusion notation $H_A \subset H$.

Definition 2.17. Let H be a Hilbert space, A be a densely defined linear operator which is strongly positive, and $f \in H$ be a given vector. Then $u \in H_A$ is called the weak solution of the operator equation $Au = f$ if

$$\langle u, v \rangle_A = \langle f, v \rangle \qquad \forall\, v \in H_A.$$

Proposition 2.18. *If the operator A satisfies the assumptions in Definition 2.17, then for every $f \in H$ there exists a unique weak solution of equation $Au = f$.*

Thus the energy space of a differential operator plays a fundamental role in the weak solution of boundary value problems. In connection with this, it also plays a key role when one looks for a self-adjoint and surjective extension $\widetilde{A}$ of A, called the Friedrichs extension, whose domain satisfies $D(A) \subset D(\widetilde{A}) \subset H_A$. For more details about unbounded operators and operator extensions we refer to the books [15, 50, 64, 66].

2.1.2 Sobolev spaces

Here we go through the definitions and the main properties of Sobolev spaces. These function spaces play the role of the abstract Hilbert space when the weak solution of a differential equation is considered. We mainly follow the treatment given in [5, 66]. Some theorems are stated in a simplified form, further details can be found in the aforementioned books or in the classical monograph [1].

Definition 2.19. Let $\Omega \subset \mathbb{R}^d$ be a bounded domain, and let V be a function space in $\mathbb{R}^{d-1}$. The boundary $\partial\Omega$ is of class V if for each point $x_0 \in \partial\Omega$ there exists an $r > 0$ and a function $\varphi \in V$ such that (after the transformation of the coordinate system, if necessary) we have

$$\Omega \cap B(x_0, r) = \{x \in B(x_0, r) :\ x_d > \varphi(x_1, \ldots, x_{d-1})\}.$$

In particular, when V is the class of Lipschitz continuous functions, then we say Ω is a Lipschitz domain. When $V = C^k(\Omega)$, then we say Ω is a C^k domain.

Definition 2.20. Let $\Omega \subset \mathbb{R}^n$ be a bounded, Lebesgue-measurable domain, $1 \leq p \leq \infty$. For a Lebesgue-measurable function $f : \Omega \to \mathbb{R}$ define the p-norm

$$\|f\|_{L^p(\Omega)} = \begin{cases} \left(\displaystyle\int_\Omega |f|^p \right)^{1/p} & \text{if } 1 \leq p < \infty, \\ \inf \left\{ \displaystyle\sup_{\Omega \setminus N} |f| : N \subset \Omega \text{ has measure zero} \right\} & \text{if } p = \infty. \end{cases}$$

The space $L^p(\Omega)$ consists of those functions whose the p-norm is finite.

Definition 2.21. Let $k \in \mathbb{N}$, $1 \leq p \leq \infty$. The Sobolev space $W^{k,p}(\Omega)$ consists of those functions $u \in L^p(\Omega)$ such that for each multi-index α the distributional derivatives $\partial^\alpha u$ exist up to order k, and $\partial^\alpha u \in L^p(\Omega)$. The norm is defined as

$$\|u\|_{W^{k,p}(\Omega)} = \begin{cases} \left(\displaystyle\sum_{|\alpha| \leq k} \|\partial^\alpha u\|^p_{L^p(\Omega)} \right)^{1/p} & \text{if } 1 \leq p < \infty, \\ \displaystyle\max_{|\alpha| \leq k} \|\partial^\alpha u\|_{L^\infty(\Omega)} & \text{if } p = \infty. \end{cases}$$

When $p = 2$, $W^{k,2}(\Omega)$ is denoted by $H^k(\Omega)$. We also introduce a seminorm on the space $H^k(\Omega)$ as

$$|u|_{H^k(\Omega)} = \left(\sum_{|\alpha| = k} \|\partial^\alpha u\|^2_{L^2(\Omega)} \right)^{1/2}.$$

Theorem 2.22. *The Sobolev space $W^{k,p}(\Omega)$ is a Banach space, and $H^k(\Omega)$ is a Hilbert space with the inner product*

$$\langle u, v \rangle_{H^k(\Omega)} = \int_\Omega \sum_{|\alpha| \leq k} \partial^\alpha u \; \partial^\alpha v.$$

Definition 2.23. Let $W^{k,p}_0(\Omega)$ be the closure of $C^\infty_0(\Omega)$ in the $\|\cdot\|_{W^{k,p}(\Omega)}$ norm. For $p = 2$ we write $H^k_0(\Omega)$.

The spaces $H^1(\Omega)$ and $H^1_0(\Omega)$ are of particular importance in the theory of second order elliptic PDEs. The latter is a closed subspace of $H^1(\Omega)$, thus it is a Hilbert space with the inherited $\|\cdot\|_{H^1(\Omega)}$ norm. But it is also a Hilbert space with the $|\cdot|_{H^1(\Omega)}$ seminorm. The proof relies on the following well-known result.

Theorem 2.24 (Poincaré–Friedrichs inequality)**.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain,*

then there exists a constant $\nu > 0$ depending only on Ω such that

$$\nu \|u\|_{L^2(\Omega)} \leq \|\nabla u\|_{L^2(\Omega)} \qquad \forall\ u \in H_0^1(\Omega). \tag{2.2}$$

Corollary 2.25. *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain, then the norms $\|\cdot\|_{H^1(\Omega)}$ and $|\cdot|_{H^1(\Omega)}$ are equivalent on $H_0^1(\Omega)$.*

Theorem 2.26 (Rellich)**.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain,*

1. *then the embedding $H_0^1(\Omega) \hookrightarrow L^2(\Omega)$ is compact;*
2. *moreover, if Ω is a Lipschitz domain, then the embedding $H^1(\Omega) \hookrightarrow L^2(\Omega)$ is compact.*

Theorem 2.27. *Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain. Then there exists a unique continuous linear operator $\gamma : H^1(\Omega) \to L^2(\partial\Omega)$ such that $\gamma u = u\big|_{\partial\Omega}$ for any $u \in H^1(\Omega) \cap C(\overline{\Omega})$. The operator γ is compact.*

This mapping is called the trace operator and for $u \in H^1(\Omega)$ the function γu is called the generalized boundary value of u. We will not go into further details, but it is worth mentioning that the trace operator, as a mapping from $H^1(\Omega)$ to $H^{1/2}(\partial\Omega)$ is surjective. For fractional Sobolev spaces and spaces over boundaries we refer to the books [1, 5].

2.2 Generalized conjugate gradient methods

Let us consider the equation

$$Au = b \tag{2.3}$$

in H, where H is a Hilbert space, $A : H \to H$ is a linear operator and $b \in H$ is a given vector. In order to ensure the well-posedness of (2.3), assume that A has a bounded inverse. When H is finite dimensional, e.g. $H = \mathbb{R}^n$ then (2.3) is a linear algebraic system.

The generalized conjugate gradient, least square (abbreviated as GCG-LS) method is constructed as follows, see in [6, 7]. There are two types of the GCG-LS algorithm: the full and the so-called truncated versions. The definition also involves an integer $s \in \mathbb{N}$, further, we let $s_k = \min\{k, s\}$, $(k \geq 0)$. The full version uses all the previous search directions to construct the sequence of approximate solutions (u_k) and search directions (d_k) such that the vectors Ad_k are linearly independent and u_k minimizes the residual norm corresponding to (2.3) in the subspace spanned by the first k search directions. The truncated version of the algorithm uses only the previous $s+1$ directions (GCG-LS(s) for short). The GCG-LS(s) algorithm is as follows:

Algorithm 2.28 (GCG–LS(s))**.**

- Let $u_0 \in H$ be arbitrary and let $r_0 = Au_0 - b$, $d_0 = -r_0$;
- For any $k \in \mathbb{N}$, when $u_k,\ d_k,\ r_k$ are obtained, let
  - the numbers $\alpha_{k-j}^{(k)} \quad (j = 0, \dots, k)$ be the solution of the system

$$\sum_{j=0}^{s_k} \alpha_{k-j}^{(k)} \langle Ad_{k-j}, Ad_{k-l} \rangle = -\langle r_k, Ad_{k-l} \rangle \quad (0 \le l \le s_k)$$

  - $u_{k+1} = u_k + \sum_{j=0}^{s_k} \alpha_{k-j}^{(k)} d_{k-j}$;
  - $r_{k+1} = r_k + \sum_{j=0}^{s_k} \alpha_{k-j}^{(k)} Ad_{k-j}$;
  - $\beta_{k-j}^{(k)} = \dfrac{\langle Ar_{k+1}, Ad_{k-j} \rangle}{\|Ad_{k-j}\|^2} \quad (j = 0, \dots, s_k)$;
  - $d_{k+1} = -r_{k+1} + \sum_{j=0}^{s_k} \beta_{k-j}^{(k)} d_{k-j}$.

The full version (called GCG-LS method) can be obtained by setting formally $s = +\infty$, whilst for finite s we get the truncated GCG-LS(s) algorithm. An interesting case arises when $s = 0$, since it requires only the current search direction, which property makes it computationally favourable.

Algorithm 2.29 (GCG–LS(0))**.**

- Let $u_0 \in H$ be arbitrary, $r_0 = Au_0 - b$, $d_0 = -r_0$;
- For given $u_k,\ d_k$, and residual $r_k = Au_k - b$, we let
  - $u_{k+1} = u_k + \alpha_k d_k$, where $\alpha_k = -\dfrac{\langle r_k, Ad_k \rangle}{\|Ad_k\|^2}$,
  - $r_{k+1} = r_k + \alpha_k Ad_k$,
  - $d_{k+1} = -r_{k+1} + \beta_k d_k$, where $\beta_k = \dfrac{\langle Ar_{k+1}, Ad_k \rangle}{\|Ad_k\|^2}$.

The following result (cf. [6, Thm. 4.1]) states the coincidence of the two algorithms in the finite dimensional case when A^* is a polynomial of A (which holds for normal matrices), see also [20].

Theorem 2.30. *Let A be a matrix satisfying $A + A^* > 0$. Assume that there exists a real polynomial $p_m \in \mathbb{R}[x]$ of degree m such that $A^* = p_m(A)$. If $s \ge m - 1$, then the truncated GCG-LS(s) method coincides with the full GCG-LS algorithm.*

Let us turn to the convergence results. Suppose that

$$A + A^* > 0, \tag{2.4}$$

that is, A is positive definite with respect to $\langle \cdot, \cdot \rangle$. The following quantities defined below will be used in the convergence theorems of the algorithms:

$$\lambda_0 \equiv \lambda_0(A) := \inf_{\|x\|=1} \langle Ax, x \rangle > 0, \qquad \Lambda \equiv \Lambda(A) := \|A\|, \tag{2.5}$$

where the norm $\|\cdot\|$ is induced by the inner product $\langle \cdot, \cdot \rangle$.

Proposition 2.31. *If* (2.4) *holds, then with the notations of* (2.5) *estimate*

$$\left(\frac{\|r_k\|}{\|r_0\|} \right)^{1/k} \leq \left(1 - \left(\frac{\lambda_0}{\Lambda} \right)^2 \right)^{1/2} \qquad (k = 1, 2, \ldots) \tag{2.6}$$

holds for the residual $r_k = Au_k - b$ of the GCG-LS(s) algorithm.

A remarkable occurrence of the GCG-LS(0) algorithm arises when A can be decomposed as

$$A = I + C, \tag{2.7}$$

where the matrix C is antisymmetric. This most often comes from symmetric part preconditioning, in which equation (2.3) is replaced by its preconditioned form $M^{-1}Au = M^{-1}b$, where M is the symmetric part of A, that is $M = (A + A^*)/2$. The preconditioned equation has the form

$$Au = b \iff M^{-1}Au = M^{-1}b \iff (I + M^{-1}N)u = M^{-1}b,$$

where $N = A - M$ is the antisymmetric part of A. In this case $A^* = 2I - A$ with respect to the M-inner product (M is spd due to (2.4)), i.e. Theorem 2.30 holds with $p_1(t) = -t + 2$. Owing to decomposition (2.7), we have the following stronger result for matrices, which provides superlinear convergence estimate in the finite dimensional case if the eigenvalues $|\lambda_1(C)| \geq |\lambda_2(C)| \geq \ldots \geq |\lambda_n(C)|$ approach zero.

Proposition 2.32. *If assumptions* (2.4) *and* (2.7) *hold, then*

$$\left(\frac{\|r_k\|}{\|r_0\|} \right)^{1/k} \leq \frac{2}{\lambda_0} \left(\frac{1}{k} \sum_{i=1}^{k} |\lambda_i(C)| \right) \qquad (k = 1, 2, \ldots, n). \tag{2.8}$$

One can also use the normal equation approach described in Subsection 1.2.2, i.e.

equation (2.3) can be replaced by

$$A^*Au = A^*b. \tag{2.9}$$

Here we can apply the standard symmetric CG method. Since A and b are replaced by A^*A and A^*b, respectively, we have to replace the residual vector for the normal equation by s_k, because we want to reserve the notation r_k for the original residual $r_k = Au_k - b$. Then we get $s_k = A^*r_k$.

Algorithm 2.33 (CGN)**.**

- Let $u_0 \in H$ be arbitrary, $r_0 = Au_0 - b,\ s_0 = d_0 = A^*r_0$;
- For given $u_k,\ d_k,\ s_k$, and $r_k = Au_k - b$, we let
  - $z_k = Ad_k$,
  - $u_{k+1} = u_k + \alpha_k d_k, \quad r_{k+1} = r_k + \alpha_k z_k,$ where $\alpha_k = -\dfrac{\langle r_k, z_k\rangle}{\|z_k\|^2}$,
  - $s_{k+1} = A^*r_{k+1}$,
  - $d_{k+1} = s_{k+1} + \beta_k d_k,$ where $\beta_k = \dfrac{\|s_{k+1}\|^2}{\|s_k\|^2}$.

The convergence estimate comes directly from the linear convergence estimate (1.30) of the symmetric CG method. Since A is replaced by A^*A, $\|e_k\|_{A^*A} = \|Ae_k\| = \|r_k\|$ and $\kappa(A^*A) = \kappa(A)^2$, the following result is obtained.

Corollary 2.34. *If* (2.4) *holds, then using the notations in* (2.5) *we have*

$$\left(\frac{\|r_k\|}{\|r_0\|}\right)^{1/k} \le 2^{1/k}\,\frac{\kappa(A)-1}{\kappa(A)+1} \le 2^{1/k}\,\frac{\Lambda-\lambda_0}{\Lambda+\lambda_0} \qquad (k = 1, 2, \ldots). \tag{2.10}$$

If the decomposition (2.7) is valid in the finite dimensional case, then by $A^*A = I + (C^* + C + C^*C)$, the superlinear convergence estimate (1.32) implies

Corollary 2.35. *If assumptions* (2.4) *and* (2.7) *hold, then*

$$\left(\frac{\|r_k\|}{\|r_0\|}\right)^{1/k} \le \frac{2}{\lambda_0^2}\left(\frac{1}{k}\sum_{i=1}^{k}\big(\,|\lambda_i(C^*+C)| + \lambda_i(C^*C)\big)\right) \qquad (k = 1, 2, \ldots, n). \tag{2.11}$$

2.3 Equivalent and compact-equivalent operators in Hilbert space

Let us consider a system of linear equations which is derived from the discretization of some elliptic differential operator. The main idea of constructing a preconditioner for the discrete system is the following: approximate the original differential operator

with another elliptic operator, which is close to the original one in some sense, and use its discretization as a preconditioning matrix. A general theory has been developed using the notion of equivalent operators, which has been introduced and investigated in the aspect of linear convergence in [21]. With the notions of Subsection 1.2.1 the main requirements are that systems with P_h should be easier to solve than with A_h and the condition number of the preconditioned matrix $P_h^{-1}A_h$ should be bounded above, where the upper bound is independent of the discretization parameter.

Following [21], we sketch the basic notions of operator equivalence, further details can be found in [23, 31, 45, 46].

Definition 2.36. Let $A, P : W \to V$ be linear operators between the Hilbert spaces W and V. The operators A and P are V-norm equivalent on a set $D \subset D(A) \cap D(P)$ if there exist $0 < \alpha \leq \beta < \infty$ such that

$$\alpha \leq \frac{\|Au\|_V}{\|Pu\|_V} \leq \beta$$

for any $u \in D$ such that the ratio is defined.

If D is sufficiently dense (that is, D is dense in $D(A)$ and $D(P)$, further $A(D)$ is dense in $R(A)$ and $P(D)$ is dense in $R(P)$), then it follows that $\kappa(AP^{-1}) \leq \beta/\alpha$, i.e. the right condition number is bounded. Similarly, for injective operators the W-norm equivalence of A^{-1} and P^{-1} implies the boundedness of $\kappa(P^{-1}A)$. The notion of equivalence can be defined between the families of operators $(A_h)_{h>0}$ and $(P_h)_{h>0}$, where the pointwise limit operators A and P exist. When the operators A_h and P_h are equivalent for any $h > 0$ and the bounds $\alpha_h \geq \alpha > 0$ and $\beta_h \leq \beta < \infty$ can be chosen independently of h, then the families (A_h) and (P_h) are called uniform V-norm equivalent. It can be shown (cf. [21, Thm. 2.12]) that the uniform V-norm equivalence of the families (A_h) and (P_h) implies the V-norm equivalence of the limit operators. The converse statement holds (cf. [21, Thm. 2.15]) if A_h and P_h are obtained via orthogonal projections from A and P, furthermore A and P are equivalent to the families (A_h) and (P_h), respectively.

The notion of operator equivalence given above is convenient when L^2-equivalence of elliptic differential operators of second order is considered. Here the uniform boundedness of the condition number of $P_h^{-1}A_h$ (or $A_hP_h^{-1}$) is ensured when P^* and A^* (or P and A) have the same boundary conditions. But it is more useful to use H^1-equivalence, i.e. equivalence based on the weak formulation of the operators, since in this case less strict regularity assumptions are needed. The main outcome of [46] is that the H^1-condition number of $P_h^{-1}A_h$ is bounded independently of h if and only if

A and P have homogeneous Dirichlet boundary conditions on the same portion of the boundary.

Here we give a uniform treatment of elliptic differential operators on the operator level using the weak formulation to handle the equivalence and compact-equivalence of the operators in a general setting. We follow the treatment given in [10].

Let H be a real Hilbert space and consider the operator equation

$$Lu = g \tag{2.12}$$

with a linear unbounded operator L in H, where $g \in H$ is given. We would like to consider its preconditioned form in weak sense in an energy space of a suitable symmetric operator.

Let $S : D(S) \subset H \to H$ be an unbounded linear symmetric operator which satisfies the coercivity property

$$\langle Su, u\rangle \geq p \|u\|^2 \qquad \forall\, u \in D(S) \tag{2.13}$$

for some $p > 0$ constant. Let $H_S \subset H$ denote the energy space of S (see Definition 2.15).

Definition 2.37. Let S be a linear symmetric coercive operator in H. A linear operator L is said to be S-bounded and S-coercive if

1. $D(L) \subset H_S$ and $D(L)$ is dense in H_S in $\|\cdot\|_S$ norm;
2. there exists $M > 0$ such that

$$|\langle Lu, v\rangle| \leq M \|u\|_S \|v\|_S \qquad \forall\, u, v \in D(L); \tag{2.14}$$

3. there exists $m > 0$ such that

$$\langle Lu, u\rangle \geq m \|u\|_S^2 \qquad \forall\, u \in D(L). \tag{2.15}$$

The set of S-bounded and S-coercive operators is denoted by $BC_S(H)$.

Definition 2.38. If $L \in BC_S(H)$ then let $L_S \in B(H_S)$ be defined by the identity

$$\langle L_S u, v\rangle_S = \langle Lu, v\rangle \qquad (u, v \in D(L)).$$

The definition makes sense, since L_S represents the unique extension of the bounded bilinear form $(u, v) \mapsto \langle Lu, v\rangle$ from $D(L)$ to H_S. Because of the density of $D(L)$ in

H_S, inequalities (2.14) and (2.15) hold in H_S for the operator L_S, i.e.

$$|\langle L_S u, v\rangle|_S \le M \|u\|_S \|v\|_S, \qquad \langle L_S u, u\rangle_S \ge m \|u\|_S^2 \qquad (u, v \in H_S). \tag{2.16}$$

Remark 2.39. If $R(L) \subset R(S)$, then the operator L_S restricted to $D(L)$ is nothing else than $S^{-1}L$.

Proposition 2.40. *(cf. [11], Prop. 3.4) Let S be a linear symmetric operator satisfying* (2.13) *and L and K be S-bounded and S-coercive operators. Then*

1. *L_S and K_S are H_S-norm equivalent;*
2. *L_S^{-1} and K_S^{-1} are H_S-norm equivalent.*

Remark 2.41. If $L \in BC_S(H)$, then L_S and the identity operator I are H_S-norm equivalent.

Definition 2.42. For a given operator $L \in BC_S(H)$, we call $u \in H_S$ the weak solution of equation (2.12) if

$$\langle L_S u, v\rangle_S = \langle g, v\rangle \qquad \forall\ v \in H_S. \tag{2.17}$$

The existence and uniqueness of the weak solution come from the Lax–Milgram lemma (cf. Theorem 2.2): the boundedness and coercivity of the bilinear form $(u, v) \mapsto \langle L_S u, v\rangle_S$ is a straightforward consequence of (2.16) and the linear functional $v \mapsto \langle g, v\rangle$ is bounded in H_S by the coercivity of S.

The theory of compact-equivalent operators has been developed in [10] and summarized in [11]. Here the compact-equivalence of the original and the preconditioning operators ensures the mesh independent superlinear convergence rate when the CGN algorithm 2.33 is used for the discretized system. In Section 2.4 similar results will be obtained for the GCG-LS algorithm 2.28 by using the compact normal operator framework.

Definition 2.43. Let L and K be S-bounded and S-coercive operators in H. We call them compact-equivalent in H_S if

$$L_S = \mu K_S + Q_S \tag{2.18}$$

for some constant $\mu > 0$ and compact operator $Q_S \in B(H_S)$.

As an important special case, we can consider compact-equivalence with $\mu = 1$ for the operators L and S as in Definition 2.37. Then

$$L_S = I + Q_S \tag{2.19}$$

with some compact operator Q_S. This comes from the fact that S itself is S-bounded and S-coercive and the corresponding operator S_S is the identity operator on H_S. This means that if the operators L and S are compact-equivalent, then L_S can be decomposed as the sum of the identity and a compact operator.

Let us consider the operator equation (2.12) where $L \in BC_S(H)$, $g \in H$ and $u \in H_S$ is the weak solution defined in (2.17). To solve it numerically, let

$$V_h = \operatorname{span}\{\varphi_1, \varphi_2, \ldots, \varphi_n\} \subset H_S$$

be a finite dimensional subspace of dimension n and

$$\mathbf{L}_h = \left\{\langle L_S\varphi_i, \varphi_j\rangle_S\right\}_{i,j=1}^n, \qquad \mathbf{g}_h = \left\{\langle g, \varphi_j\rangle\right\}_{j=1}^n. \tag{2.20}$$

Then the discrete solution $u_h \in V_h$ is $u_h = \sum_{i=1}^n c_i\varphi_i$, where $\mathbf{c} = (c_1, \ldots, c_n) \in \mathbb{R}^n$ is the solution of the linear system

$$\mathbf{L}_h\mathbf{c} = \mathbf{g}_h,$$

which is the discretized form of (2.17). Now assume that L and S are compact-equivalent with $\mu = 1$, i.e. relation (2.19) holds. If S is used as a preconditioner, then the discretized form of the operator decomposition (2.19) becomes

$$\mathbf{L}_h = \mathbf{S}_h + \mathbf{Q}_h, \tag{2.21}$$

and the corresponding preconditioned form of equation (2.17) has the form

$$\left(\mathbf{I}_h + \mathbf{S}_h^{-1}\mathbf{Q}_h\right)\mathbf{c} = \mathbf{S}_h^{-1}\mathbf{g}_h, \tag{2.22}$$

where

$$\mathbf{S}_h = \left\{\langle\varphi_i, \varphi_j\rangle_S\right\}_{i,j=1}^n, \qquad \mathbf{Q}_h = \left\{\langle Q_S\varphi_i, \varphi_j\rangle_S\right\}_{i,j=1}^n. \tag{2.23}$$

If we apply Algorithm 2.33 for equation (2.22) then Corollary 2.35 holds with $C = \mathbf{S}_h^{-1}\mathbf{Q}_h$ and $\lambda_0 = m$. It has been proved in [10, Prop. 4.1] that the eigenvalues appear in (2.11) can be estimated above by the eigenvalues of the corresponding operators, thus we have

Theorem 2.44. *(cf. [10, Thm. 4.1]) Assume that $L \in BC_S(H)$, L and S are compact-equivalent with $\mu = 1$, i.e. (2.19) holds. Then the CGN algorithm 2.33 for system (2.22) yields*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \le \frac{2}{m^2}\left(\frac{1}{k}\sum_{i=1}^k \left(|\lambda_i(Q_S^* + Q_S)| + \lambda_i(Q_S^*Q_S)\right)\right) \xrightarrow{k\to\infty} 0, \tag{2.24}$$

where the right-hand side is independent of the subspace V_h.

2.4 The compact normal operator framework

Let us return to the operator equation

$$Au = b, \tag{2.25}$$

where $A : H \to H$ is a nonsymmetric linear operator on the Hilbert space H and $f \in H$ is a given vector. Assume that A has a bounded inverse to ensure the well-posedness of (2.25). Algorithms 2.28 and 2.33 can be formulated in Hilbert space without any modification. The following result, which can be found in [8, Thm. 1], is an extension of Theorem 2.30 to the infinite dimensional case.

Theorem 2.45. *Let H be a real Hilbert space and $A : H \to H$ be a bounded linear operator satisfying $A + A^* > 0$. Assume that there exists a real polynomial $p_m \in \mathbb{R}[x]$ of degree m such that $A^* = p_m(A)$. If $s \geq m - 1$, then the truncated GCG-LS(s) method coincides with the full GCG-LS algorithm.*

Remark 2.46. If there exist constants $c_1, c_2 \in \mathbb{R}$ such that $A^* = c_1 A + c_2 I$, then the truncated GCG-LS(0) method coincides with the full GCG-LS algorithm.

Let equation (2.25) have the form

$$Au \equiv (I + C)u = f \tag{2.26}$$

with a compact operator C. Denote by $\lambda_k \equiv \lambda_k(C)$ $(k \in \mathbb{N})$ the ordered eigenvalues of C, where $\lambda_k \to 0$ by the compactness of C. The superlinear convergence estimate (1.31) can be extended to the infinite dimensional case for operators that can be written in the form $A = I + C$, where C is a compact normal operator, which ensures that the eigenvectors form a complete orthonormal basis in H (cf. Remark 1.7, Theorem 2.6), proved in [63]. Using the boundedness of A^{-1}, estimate (1.31) has the following infinite dimensional counterpart (cf. [9, Thm. 2]):

Theorem 2.47. *Let H be a complex separable Hilbert space and $C : H \to H$ be a compact normal operator on H with ordered eigenvalues $\lambda_k(C)$ $(k \in \mathbb{N})$. Suppose that A can be decomposed as*

$$A = I + C, \tag{2.27}$$

where I is the identity operator and assume that A has a bounded inverse. Then the

GCG-LS algorithm 2.28 yields for all $k \in \mathbb{N}$

$$\left(\frac{\|e_k\|_A}{\|e_0\|_A}\right)^{1/k} \leq 2\left\|A^{-1}\right\|\left(\frac{1}{k}\sum_{i=1}^{k}|\lambda_i(C)|\right) \xrightarrow{k\to\infty} 0. \tag{2.28}$$

2.4.1 Preconditioned operator equations and superlinear convergence

When the conjugate gradient method is applied to systems arising from the discretization of elliptic PDEs, the spectral condition number that appears in the linear convergence estimate (1.30) tends to infinity as the mesh is refined, as pointed out in (1.20). Thus suitable preconditioning is required to obtain a mesh independent convergence bound. The application of the conjugate gradient method for the preconditioned form of (2.25) has been investigated in the aspect of linear convergence using the framework of equivalent operators in Hilbert space (cf. [21, 46]). These results have been extended in [8, 9], where superlinear convergence has been proved for operator equations and mesh independent bound for the estimate obtained for the discretized systems using the GCG-LS algorithm.

Let us consider an operator equation

$$Lu = g \tag{2.29}$$

with an unbounded linear operator $L : D \subset H \to H$ defined on a dense domain D, and with some $g \in H$. Consider a preconditioned version of (2.29) which has the form (2.27) in a suitable energy space. Equation (2.29) is assumed to satisfy the following

Assumptions 2.48. *Assume that*

(i) the operator L *is decomposed in* $L = S + Q$ *on its domain* D *where* S *is a self-adjoint operator in* H*;*

(ii) S *is a strongly positive operator, i.e. there exists* $p > 0$ *such that*

$$\langle Su, u\rangle \geq p\,\|u\|^2 \qquad \forall\ u \in D; \tag{2.30}$$

(iii) there exists $\varrho > 0$ *such that* $\operatorname{Re}\langle Lu, u\rangle \geq \varrho\,\langle Su, u\rangle \quad \forall\ u \in D;$

(iv) the operator Q *can be extended to the energy space* H_S*, and then* $S^{-1}Q$ *is assumed to be a compact normal operator on* H_S*.*

We recall that the energy space H_S is the completion of D under the energy inner product $\langle u, v\rangle_S = \langle Su, v\rangle$ $(u, v \in D)$, and the corresponding norm has the obvious notation $\|\cdot\|_S$. Condition *(ii)* implies $H_S \subset H$ (cf. Proposition 2.16). By Corollary

2.14, conditions (i)-(ii) on S in Assumptions 2.48 imply that $R(S) = H$ and hence $S^{-1}Q$ makes sense. Since now S is onto, the use of the weakly defined operators L_S ans Q_S (in contrast with Section 2.3) is not needed (cf. Remark 2.39).

Remark 2.49. If $\operatorname{Re}\langle Qu, u\rangle \geq 0$ holds for every $u \in D$, then (iii) holds with $\varrho = 1$. This is valid if Q is antisymmetric.

Remark 2.50. The normality of $S^{-1}Q$ on the space H_S means that it is S-normal, i.e. the operator $(S^{-1}Q)^*_S$, the adjoint of $S^{-1}Q$ with respect to the inner product $\langle\cdot,\cdot\rangle_S$, commutes with $S^{-1}Q$.

Now we replace equation (2.29) by its preconditioned form

$$S^{-1}Lu = f \equiv S^{-1}g. \tag{2.31}$$

Then the full GCG-LS algorithm 2.28 in H_S is as follows. Here for better algorithmization four sequences $u_k,\ d_k,\ r_k,\ z_k$ are constructed, and the notation $Ad_j = z_j$ is used throughout the algorithm (where A is replaced by $S^{-1}L$).

Algorithm 2.51 (Preconditioned GCG–LS(s))**.**

- Let $u_0 \in D$ be arbitrary, and let r_0 be the solution of $Sr_0 = Lu_0 - g$, $d_0 = -r_0$, and z_0 be the solution of $Sz_0 = Ld_0$;
- For any $k \in \mathbb{N}$, when $u_k,\ d_k,\ r_k,\ z_k$ are obtained, let
  - the numbers $\alpha^{(k)}_{k-j}\quad (j = 0, \dots, k)$ be the solution of the system

$$\sum_{j=0}^{s_k} \alpha^{(k)}_{k-j} \langle Sz_{k-j}, z_{k-l}\rangle = -\langle r_k, Sz_{k-l}\rangle \quad (0 \leq l \leq s_k)$$

  - $u_{k+1} = u_k + \sum\limits_{j=0}^{s_k} \alpha^{(k)}_{k-j} d_{k-j}$;
  - $r_{k+1} = r_k + \sum\limits_{j=0}^{s_k} \alpha^{(k)}_{k-j} z_{k-j}$;
  - $\beta^{(k)}_{k-j} = \dfrac{\langle Lr_{k+1}, z_{k-j}\rangle}{\|z_{k-j}\|^2_S} \quad (j = 0, \dots, s_k)$;
  - $d_{k+1} = -r_{k+1} + \sum\limits_{j=0}^{s_k} \beta^{(k)}_{k-j} d_{k-j}$;
  - z_{k+1} be the solution of $Sz_{k+1} = Ld_{k+1}$.

In the truncated GCG-LS(0) algorithm 2.29 the vectors z_k can be determined within the kth cycle since no previous indices are used:

Algorithm 2.52 (Preconditioned GCG–LS(0))**.**

- Let $u_0 \in D$ be arbitrary, and let r_0 be the solution of $Sr_0 = Lu_0 - g$, $d_0 = -r_0$;
- For any $k \in \mathbb{N}$, when $u_k,\ d_k,\ r_k$ are obtained, let
  - z_k be the solution of $Sz_k = Ld_k$;
  - $u_{k+1} = u_k + \alpha_k d_k,$ where $\alpha_k = -\dfrac{\langle r_k, Sz_k\rangle}{\langle Sz_k, z_k\rangle}$,
  - $r_{k+1} = r_k + \alpha_k z_k$,
  - $d_{k+1} = -r_{k+1} + \beta_k d_k,$ where $\beta_k = \dfrac{\langle Lr_{k+1}, z_k\rangle}{\langle Sz_k, z_k\rangle}$.

Owing to the decomposition of L, equation (2.31) is equivalent to

$$\left(I + S^{-1}Q\right) u = f \equiv S^{-1}g, \tag{2.32}$$

that is, it has the form (2.27) with

$$A = I + S^{-1}Q.$$

Using Assumptions 2.48 it has been shown in [9] that A in a linear operator in H_S which has a bounded inverse, hence equation (2.32) has a unique solution $u \in H_S$. This can be considered as the weak solution of (2.29), since

$$\langle u, v\rangle_S + \langle Qu, v\rangle = \langle g, v\rangle \qquad \forall\ v \in H_S. \tag{2.33}$$

Furthermore, condition (*iii*) implies that $\|u\|_A = \|u\|_L$ and $\|A^{-1}\|_S \le 1/\varrho$ holds. Then we have

Theorem 2.53. *(cf. [9, Thm. 3]) Let Assumptions 2.48 hold. Then the GCG-LS algorithm 2.28 applied for equation* (2.31) *in H_S yields for all $k \in \mathbb{N}$*

$$\left(\frac{\|e_k\|_L}{\|e_0\|_L}\right)^{1/k} \le \frac{2}{\varrho}\left(\frac{1}{k}\sum_{i=1}^{k}\left|\lambda_i\left(S^{-1}Q\right)\right|\right) \xrightarrow{k\to\infty} 0, \tag{2.34}$$

where $\lambda_k\left(S^{-1}Q\right)$ $(k \in \mathbb{N})$ are the ordered eigenvalues of the compact normal operator $S^{-1}Q$.

Equation (2.29) can be solved numerically using Galerkin discretization. Let

$$V_h = \operatorname{span}\{\varphi_1, \varphi_2, \ldots, \varphi_n\} \subset H_S$$

be a finite dimensional subspace of dimension n. Then the discrete solution $u_h \in V_h$ is $u_h = \sum_{i=1}^{n} c_i \varphi_i$, where $\mathbf{c} = (c_1, c_2, \ldots, c_n) \in \mathbb{R}^n$ is the solution of the linear algebraic system

$$\mathbf{L}_h \mathbf{c} = \mathbf{g}_h, \tag{2.35}$$

where $\mathbf{g}_h = \{\langle g, \varphi_j \rangle\}_{j=1}^{n}$ and the matrix $\mathbf{L}_h$ is defined as $\mathbf{L}_h = \mathbf{S}_h + \mathbf{Q}_h$, where

$$\mathbf{S}_h = \left\{\langle \varphi_i, \varphi_j \rangle_S\right\}_{i,j=1}^{n}, \qquad \mathbf{Q}_h = \left\{\langle Q\varphi_i, \varphi_j \rangle\right\}_{i,j=1}^{n}.$$

Hence equation (2.35) can be written as

$$\left(\mathbf{S}_h + \mathbf{Q}_h\right) \mathbf{c} = \mathbf{g}_h,$$

which is the discretized form of (2.33). If the operator S is used as a preconditioner, then the discretized form of the preconditioned operator equation (2.32) becomes

$$\left(\mathbf{I}_h + \mathbf{S}_h^{-1}\mathbf{Q}_h\right) \mathbf{c} = \mathbf{S}_h^{-1}\mathbf{g}_h. \tag{2.36}$$

Similarly to the mesh independent result of CGN algorithm in the previous section, the eigenvalues of the matrix in the finite dimensional estimate (2.8) can be estimated above by the eigenvalues of the corresponding operator, thus we have

Theorem 2.54. *(cf. [9, Cor. 4]) Suppose that H is a complex separable Hilbert space, Assumptions 2.48 are satisfied and the matrix $\mathbf{S}_h^{-1}\mathbf{Q}_h$ is $\mathbf{S}_h$-normal. Then the GCG-LS algorithm 2.28 for system* (2.36) *yields*

$$\left(\frac{\|e_k\|_{\mathbf{L}_h}}{\|e_0\|_{\mathbf{L}_h}}\right)^{1/k} \leq \frac{2}{\varrho}\left(\frac{1}{k}\sum_{i=1}^{k}\left|\lambda_i(S^{-1}Q)\right|\right) \xrightarrow{k\to\infty} 0, \tag{2.37}$$

where the right-hand side is independent of the subspace V_h.

2.4.2 Symmetric part preconditioning

Here the symmetric part preconditioning strategy is summarized briefly, i.e. when the symmetric part of an operator is used as preconditioning operator. It has been introduced and analysed in [14, 62] (see also [6]), and efficiently applied to nonsymmetric elliptic problems (convection-diffusion equations). For the solution of discretized elliptic problems it has proved an efficient tool, see in [8, 9] for problems with Dirichlet boundary conditions, and in [34] for mixed problems.

Strong symmetric part

Consider equation (2.29) with the additional coercivity assumption

$$\operatorname{Re}\langle Lu, u\rangle \geq p\,\|u\|^2 \qquad \forall\, u \in D(L) \tag{2.38}$$

with some positive constant $p > 0$. Let S and Q be the symmetric and antisymmetric part of L, that is

$$Su = \frac{Lu + L^*u}{2}, \qquad Qu = \frac{Lu - L^*u}{2} \qquad \forall\, u \in D(L) \cap D(L^*)$$

and assume that $D(L) \cap D(L^*)$ is dense in H, $R(S) = H$, further, Q can be extended to H_S and $S^{-1}Q$ is a compact operator on H_S. Then S is self-adjoint by Corollary 2.14 and L is decomposed as $L = S + Q$ on the dense domain $D := D(L) \cap D(L)^*$. Since $\langle Su, u\rangle = \operatorname{Re}\langle Lu, u\rangle$, S is strongly positive by (2.38). The operator $S^{-1}Q$ is normal in H_S, since

$$\left\langle S^{-1}Qu, v\right\rangle_S = \langle Qu, v\rangle = -\langle u, Qv\rangle = -\left\langle u, S^{-1}Qv\right\rangle_S \qquad \forall\, u \in H_S,$$

which means that $S^{-1}Q$ in H_S inherits the antisymmetry of Q in D.

Thus we have proved that using symmetric part preconditioning Assumptions 2.48 are satisfied and Theorem 2.53 holds (with $\varrho = 1$) for equation (2.31) with the symmetric part S of L as preconditioner. Moreover, the antisymmetry of $S^{-1}Q$ implies that $A_S^* = 2I - A$ (see [8] and the analogous argument for matrices on page 25), thus the truncated GCG-LS(0) algorithm coincides with the full version.

Remark 2.55. It follows from the above argument that if $\mathbf{S}_h$ is the symmetric part of $\mathbf{L}_h$ in the preconditioned form of the discretized equation (2.36), then estimate (2.37) holds for the GCG-LS(0) algorithm with $\varrho = 1$, the error is measured in $\mathbf{S}_h$-norm and the $\mathbf{S}_h$-normality of $\mathbf{S}_h^{-1}\mathbf{Q}_h$ does not need to be assumed, since it is automatically satisfied.

Weak symmetric part

When $D \subset H$ is not known to be dense, the symmetric part of L has to be defined in weak sense. We go through the basic steps of the construction, further details and the proofs can be found in [34]. Assume that (2.38) hold and let the weak symmetric part of L be the sesquilinear form

$$\langle u, v\rangle_S = \frac{\langle Lu, v\rangle + \langle u, Lv\rangle}{2} \qquad \forall\, u, v \in D(L),$$

which defines an inner product on $D(L)$. Then H_S is defined as the completion of $D(L)$ under the inner product $\langle \cdot, \cdot \rangle_S$. Assume further that there exists $M > 0$ such that

$$|\langle Lu, v \rangle| \leq M \|u\|_S \|v\|_S \qquad \forall\, u, v \in D(L). \tag{2.39}$$

Then $S^{-1}Q$ can be replaced by the operator $Q_S : H_S \to H_S$, defined as

$$\langle Q_S u, v \rangle_S = \frac{\langle u, v \rangle_L - \overline{\langle v, u \rangle_L}}{2} \qquad \forall\, u, v \in H_S, \tag{2.40}$$

where the bounded sesquilinear form $\langle \cdot, \cdot \rangle_L$ is the unique extension of the form $(u, v) \mapsto \langle Lu, v \rangle$ from $D(L)$ to H_S. Then we have

$$\langle u, v \rangle_L = \langle u, v \rangle_S + \langle Q_S u, v \rangle_S \qquad \forall\, u, v \in H_S.$$

It has been shown that there exists $f \in H_S$ such that $\langle g, v \rangle = \langle f, v \rangle_S$ for any $v \in H_S$, thus the weak form

$$\langle u, v \rangle_L = \langle g, v \rangle \qquad \forall\, v \in H_S$$

of (2.29) becomes

$$(I + Q_S)\, u = f$$

in H_S, which is a generalized form of (2.32). Assuming that (2.38) and (2.39) hold and Q_S is compact on H_S, it has been proved in [34] that the conditions of Theorem 2.47 satisfied with $A = I + Q_S$ in H_S. Thus estimate (2.28) holds when $\lambda_i(S^{-1}Q)$ is replaced by $\lambda_i(Q_S)$ with $\|A^{-1}\|_S \leq 1$, the error is measured in $\|\cdot\|_S$ norm and the truncated algorithm coincides with the full version.

Remark 2.56. The construction of the weak symmetric part and the weak form of the equation is analogous to the construction of the weakly defined operator L_S in H_S and the weak solution (2.17). The main difference is that in Definitions 2.37-2.38 the operator S was given in advance, but here it was constructed directly from the operator L and only in weak sense.

Remark 2.57. Since now the weak form of the operators are used, the matrices $\mathbf{L}_h$, $\mathbf{S}_h$ and $\mathbf{Q}_h$ are defined as in (2.20) and (2.23). The observations in Remark 2.55 hold with replacing the eigenvalues of the strongly defined operator $S^{-1}Q$ by the eigenvalues of the weakly defined operator Q_S.

3. SYMMETRIC PRECONDITIONING FOR LINEAR ELLIPTIC EQUATIONS

The Hilbert space setting of the FEM enables us to estimate the superlinear convergence factors in the discrete case from above by the analogous factors in the operator level, where the latter is based on the eigenvalues of the preconditioned operator. Using the theoretical background of Section 2.4 for problems with homogeneous mixed boundary conditions, first we investigate the relation between the known theoretical convergence estimate and the numerical results (cf. [41, 42]). Then we extend the theory to the case of nonhomogeneous mixed boundary conditions using operator pairs (see [40]) and the background of Section 2.3. For FDM discretizations we do not have such abstract Hilbert space background, hence no general results exist for the mesh independent convergence of the discretized systems. The study of a special model problem is considered in Section 3.3, based on [38]. From now on, the content of the chapters consists of the author's results, published in the mentioned and other papers.

3.1 Equations with homogeneous mixed boundary conditions

In this section convection-diffusion equations are considered with the aim of investigating the relation between the theoretical convergence estimates (2.34)-(2.37) and the numerical results. The main goal of this section is twofold: first to confirm the mesh independent superlinear convergence property of the CGM when symmetric part preconditioning is applied to the FEM discretization of the boundary value problem (3.3). Second, we have also analysed cases not covered by theory through experiments, i.e. when another symmetric operator is used as a preconditioner, not only the symmetric part of the operator.

For a given densely defined operator L the standard way of constructing its symmetric part is

$$S = \frac{L + L^*}{2},$$

as described in Subsection 2.4.2. This is feasible for Dirichlet problems, but for mixed problems it is generally impossible, because the domain of L may differ from the domain of its adjoint, i. e. $D(L) \neq D(L^*)$. Hence the density property of the domain of S may

not be valid anymore, thus the definition of S requires a more general approach, it can be defined in weak sense as at the end of Subsection 2.4.2.

We would like to use Theorem 2.53 – and mainly its discrete counterpart, Theorem 2.54 – for the equation

$$Lu = g, \tag{3.1}$$

where L is a densely defined unbounded linear operator, $g \in H$ is a given vector, and L is decomposed in $L = S + Q$ on the domain $D(L)$, where S is a self-adjoint operator. Preconditioning with the operator S, we can replace equation (3.1) by its preconditioned form

$$S^{-1}Lu = S^{-1}(S+Q)u = \underbrace{(I + S^{-1}Q)}_{A}u = f \equiv S^{-1}g. \tag{3.2}$$

Based on Subsection 2.4.2, in the case of symmetric part preconditioning the S-adjoint of A is a linear polynomial of A (see Remark 2.46), thus only the truncated GCG-LS(0) algorithm will be considered. This method is closely related to the so-called CGW-method, see [14, 62]. In what follows, we summarize the application of the above theory to convection-diffusion equations, including the construction of the weak symmetric part. These results can be found in detail in [9, 34] for Dirichlet and mixed boundary conditions, respectively.

3.1.1 The problem and the algorithm in Sobolev space

In this subsection we define the linear elliptic second-order differential operator L, where the role of the abstract Hilbert space H is played by the function space $L^2(\Omega)$. Let us consider an elliptic convection-diffusion equation with mixed boundary conditions

$$\left.\begin{aligned} Lu \equiv -\Delta u + \mathbf{b}\cdot\nabla u + cu &= g \\ u\big|_{\Gamma_D} = 0, \quad \frac{\partial u}{\partial \nu} + \alpha u\big|_{\Gamma_N} &= 0 \end{aligned}\right\} \tag{3.3}$$

satisfying the following assumptions:

Assumptions 3.1. *Suppose that*

(i) $\Omega \subset \mathbb{R}^d$ *is a bounded piecewise* C^1 *domain;* Γ_D, Γ_N *are disjoint open measurable subparts of* $\partial\Omega$ *such that* $\partial\Omega = \overline{\Gamma}_D \cup \overline{\Gamma}_N$*;*

(ii) $\mathbf{b} \in C^1(\overline{\Omega})^d$, $c \in L^\infty(\Omega)$, $\alpha \in L^\infty(\Gamma_N)$ *and* $c, \alpha \geq 0$*;*

(iii) we have the coercivity properties

$$\hat{c} := c - \frac{1}{2}\operatorname{div}\mathbf{b} \geq 0 \text{ in } \Omega, \qquad \hat{\alpha} := \alpha + \frac{1}{2}(\mathbf{b}\cdot\nu) \geq 0 \text{ on } \Gamma_N; \tag{3.4}$$

(iv) $g \in L^2(\Omega)$;

(v) either $\Gamma_D \neq \emptyset$, *or* $\hat{c}$ *or* $\hat{\alpha}$ *is not constant zero.*

Let us consider the complex Hilbert space $H = L^2(\Omega)$ with the usual inner product

$$\langle u, v \rangle_{L^2(\Omega)} = \int_\Omega u\overline{v} \, \mathrm{d}x$$

and define the operator L as

$$Lu \equiv -\Delta u + \mathbf{b} \cdot \nabla u + cu$$

with the domain

$$D \equiv D(L) := \left\{ u \in H^2(\Omega) : \; u\big|_{\Gamma_D} = 0, \; \frac{\partial u}{\partial \nu} + \alpha u\big|_{\Gamma_N} = 0 \right\}, \tag{3.5}$$

which is dense in H. We have

$$\langle Lu, v \rangle_{L^2(\Omega)} = \int_\Omega \left(\nabla u \cdot \nabla \overline{v} + (\mathbf{b} \cdot \nabla u)\, \overline{v} + cu\overline{v} \right) \, \mathrm{d}x + \int_{\Gamma_N} \alpha u \overline{v} \, \mathrm{d}\sigma \quad (u, v \in D(L)). \tag{3.6}$$

The weak symmetric part of L is constructed in Subsection 2.4.2, which is now the following sesquilinear form:

$$\begin{aligned} \langle u, v \rangle_S &= \frac{1}{2} \big(\langle Lu, v \rangle_{L^2(\Omega)} + \langle u, Lv \rangle_{L^2(\Omega)} \big) \\ &= \int_\Omega (\nabla u \cdot \nabla \overline{v} + \hat{c} u \overline{v}) \, \mathrm{d}x + \int_{\Gamma_N} \hat{\alpha} u \overline{v} \, \mathrm{d}\sigma \quad (u, v \in D(L)), \end{aligned} \tag{3.7}$$

and the energy space H_S – which is defined as the completion of D under the inner product $\langle \cdot, \cdot \rangle_S$ – is

$$H_S = H^1_D(\Omega) := \left\{ u \in H^1(\Omega) : u\big|_{\Gamma_D} = 0 \right\}. \tag{3.8}$$

By (2.40) we define the operator $Q_S : H_S \to H_S$, which has the form

$$\langle Q_S u, v \rangle_S = \frac{1}{2} \left(\int_\Omega (\mathbf{b} \cdot \nabla u)\, \overline{v} \, \mathrm{d}x - \int_\Omega u \, (\mathbf{b} \cdot \nabla \overline{v}) \, \mathrm{d}x \right). \tag{3.9}$$

Here the strong form of the operator S corresponding to the sesquilinear form (3.7) cannot be used, since it is generally not known to be surjective (i.e. it may not be self-adjoint) due to the lack of H^2-regularity result on the weak solution in the presence of mixed boundary conditions. Therefore S^{-1} may not make sense, thus the precondi-

tioned GCG-LS(0) algorithm 2.52 has to be reformulated using the weak formulation of S.

Algorithm 3.2 (Preconditioned GCG–LS(0) in weak form)**.**

- Let $u_0 \in H^1_D(\Omega)$ be arbitrary, and let $r_0 \in H^1_D(\Omega)$ be the weak solution of

$$\begin{cases} -\Delta r_0 + \hat{c} r_0 = L u_0 - g \\ r_0\big|_{\Gamma_D} = 0, \quad \dfrac{\partial r_0}{\partial \nu} + \hat{\alpha} r_0\big|_{\Gamma_N} = 0; \end{cases}$$

  $d_0 = -r_0$;

- For any $k \in \mathbb{N}$, when $u_k,\ d_k,\ r_k$ are obtained, let
  - $z_k \in H^1_D(\Omega)$ be the weak solution of

$$\begin{cases} -\Delta z_k + \hat{c} z_k = L d_k \\ z_k\big|_{\Gamma_D} = 0, \quad \dfrac{\partial z_k}{\partial \nu} + \hat{\alpha} z_k\big|_{\Gamma_N} = 0, \end{cases}$$

  - $u_{k+1} = u_k + \alpha_k d_k$, where $\alpha_k = -\dfrac{\langle r_k, z_k \rangle_S}{\langle z_k, z_k \rangle_S}$,
  - $r_{k+1} = r_k + \alpha_k z_k$,
  - $d_{k+1} = -r_{k+1} + \beta_k d_k$, where $\beta_k = \dfrac{\langle r_{k+1}, z_k \rangle_L}{\langle z_k, z_k \rangle_S}$.

The following theorem shows that the assumptions on the differential equation (3.3) ensure that the assumptions on the abstract operator equation (2.29) hold.

Theorem 3.3. *Let problem* (3.3) *satisfy Assumptions 3.1. Then the PCG Algorithm 3.2 converges superlinearly, i.e. for all* $k \in \mathbb{N}$

$$\left(\frac{\|e_k\|_S}{\|e_0\|_S} \right)^{1/k} \leq \frac{2}{k} \sum_{i=1}^{k} |\lambda_i(Q_S)| \xrightarrow{k \to \infty} 0, \tag{3.10}$$

where $\lambda_i(Q_S)$ *are the ordered eigenvalues of the operator* Q_S*.*

The proof can be found in [9, Cor. 1] for Dirichlet problems and in [34, Thm. 4.1] for mixed problems.

3.1.2 FEM discretization and mesh independence

Now we consider finite element discretizations of problem (3.3). Let H_S be defined as in (3.8) and let $V_h = \operatorname{span}\{\varphi_1, \varphi_2, \ldots, \varphi_n\} \subset H_S$ be a given FEM subspace. The FEM solution $u_h \in V_h$ of equation (3.3) in V_h is $u_h = \sum_{i=1}^{n} c_i \varphi_i$, where

$\mathbf{c} = (c_1, c_2 \dots, c_n) \in \mathbb{C}^n$ is the solution of the $n \times n$ system

$$\mathbf{L}_h \mathbf{c} = \mathbf{g}_h, \tag{3.11}$$

where

$$(\mathbf{L}_h)_{i,j} = \int_\Omega \left(\nabla\varphi_i \nabla\overline{\varphi}_j + (\mathbf{b} \cdot \nabla\varphi_j)\, \overline{\varphi}_i + c\varphi_i \overline{\varphi}_j \right) \, \mathrm{d}x + \int_{\Gamma_N} \alpha\varphi_i \overline{\varphi}_j \, \mathrm{d}\sigma$$

and

$$(\mathbf{g}_h)_j = \int_\Omega g\overline{\varphi}_j.$$

Let $\mathbf{S}_h$ and $\mathbf{Q}_h$ be the symmetric and antisymmetric parts of $\mathbf{L}_h$, that is

$$\mathbf{S}_h = \frac{\mathbf{L}_h + \mathbf{L}_h^*}{2}, \qquad \mathbf{Q}_h = \mathbf{L}_h - \mathbf{S}_h.$$

Using the symmetric part $\mathbf{S}_h$ as preconditioner, equation (3.11) is replaced by

$$\mathbf{S}_h^{-1} \mathbf{L}_h \mathbf{c} = \left(\mathbf{I}_h + \mathbf{S}_h^{-1} \mathbf{Q}_h \right) \mathbf{c} = \mathbf{S}_h^{-1} \mathbf{g}_h. \tag{3.12}$$

Since Theorem 3.3 holds by Assumptions 3.1, the general mesh independent result of Theorem 2.54 and Remarks 2.55 and 2.57 imply the following

Corollary 3.4. *Let problem* (3.3) *satisfy Assumptions 3.1. Then algorithm* (2.29) *applied for* (3.12) *yields*

$$\left(\frac{\|e_k\|_{\mathbf{S}_h}}{\|e_0\|_{\mathbf{S}_h}} \right)^{1/k} \leq \frac{2}{k} \sum_{i=1}^{k} |\lambda_i(Q_S)| \xrightarrow{k\to\infty} 0, \tag{3.13}$$

where $e_k = u_k - u_h$ is the error vector, $\lambda_i(Q_S)$ are the ordered eigenvalues of the operator Q_S, hence the sequence on the right-hand side is independent of the subspace V_h and tends to zero.

3.1.3 Numerical experiments

The numerical superlinear convergence will be investigated for a special test problem. Symmetric part preconditioning is in focus to confirm the previously cited mesh independent theoretical estimate and what is more, much better results are shown than the rather pessimistic estimate (3.13). Furthermore, similar numerical results are obtained when not the symmetric part of the operator L but another symmetric

elliptic operator is used as preconditioner. The theory does not cover this case, but the numerical results show much similar behaviour.

Our test problems are the following elliptic convection-diffusion equation with two possible boundary conditions (a) and (b):

$$\begin{cases} Lu \equiv -\Delta u + \dfrac{\partial u}{\partial x} + cu = g \\ (a) \quad u\big|_{\partial\Omega} = 0 \\ (b) \quad u\big|_{\Gamma_D} = 0, \quad \dfrac{\partial u}{\partial \nu} = 0. \end{cases} \tag{3.14}$$

This special model problem has the following properties:

(i) $\Omega = [0,1] \times [0,1]$ is the unit square. We have the boundary portions

(a) $\Gamma_D = \partial\Omega$;

(b) $\Gamma_D = \{(x,y) \in \partial\Omega : \ x = 0 \text{ or } x = 1\}$;

(ii) $\mathbf{b} = (1,0)$, $c \geq 0$ is a constant;

(iii) g is a polynomial.

It is easy to verify that Assumptions 3.1 for the general problem are satisfied.

Numerous experiments have been performed in connection with the test problems. The vector $\mathbf{b} = (1,0)$ is fixed, but in the last part of this subsection convection-dominated equations will be considered, in that case $\mathbf{b} = (\eta, 0)$, $\eta \gg 1$. Denoting by c_L the constant $c \geq 0$ in operator L and c_S in operator S, the focus is on the superlinear convergence property of the two test problems. If the symmetric part of L is used for preconditioning, then $c_L = c_S$ (since $\hat{c} = c \equiv c_L$). The case of different constants in the operator L and S is also investigated.

The following notation will be used throughout this subsection for the quotient of the error vectors according to the left-hand side of estimate (3.13) in Corollary 3.4:

$$q_k := \frac{\|e_k\|_{\mathbf{S}_h}}{\|e_0\|_{\mathbf{S}_h}}, \qquad Q_k := \left(\frac{\|e_k\|_{\mathbf{S}_h}}{\|e_0\|_{\mathbf{S}_h}} \right)^{1/k}.$$

Experiment 1 In the first set of experiments equation (3.14) has been considered with boundary conditions (a) and (b), $c_L = c_S = 1$.

For $h = 1/4$ the results are much better compared with the others, which must have been caused by the very few points on the grid. The numbers in each column tend to zero which shows the superlinear convergence for every mesh parameter h. Considering

the rows, the numbers increase but the growth rate becomes slower, which is enough for a numerical evidence of the mesh independence.

Tab. 3.1: Values of Q_k, boundary conditions (a).

	1/h						
Itr.	4	8	16	32	64	128	256
1	0.06127	0.07448	0.07802	0.07892	0.07914	0.07920	0.07921
2	0.04978	0.06510	0.06904	0.07004	0.07029	0.07035	0.07037
3	0.03809	0.05820	0.06291	0.06410	0.06440	0.06447	0.06449
4	0.03332	0.05195	0.05761	0.05903	0.05939	0.05948	0.05950
5	0.02904	0.04618	0.05277	0.05443	0.05485	0.05495	0.05498
6	0.02555	0.04156	0.04843	0.05034	0.05082	0.05094	0.05097
7	0.01888	0.03957	0.04461	0.04671	0.04726	0.04739	0.04743
8	0.01778	0.03922	0.04148	0.04352	0.04412	0.04427	0.04431
9	0.01958	0.03784	0.03981	0.04073	0.04135	0.04173	0.04377

The results for the mixed problem in Table 3.2 are similar to the previous, simpler problem in Table 3.1.

Tab. 3.2: Values of Q_k, boundary conditions (b)

	1/h						
Itr.	4	8	16	32	64	128	256
1	0.08893	0.09945	0.10219	0.10289	0.10306	0.10311	0.10312
2	0.07836	0.09024	0.09317	0.09390	0.09409	0.09413	0.09414
3	0.07105	0.08428	0.08753	0.08835	0.08855	0.08860	0.08862
4	0.06726	0.07962	0.08292	0.08375	0.08397	0.08401	0.08403
5	0.06047	0.07567	0.07911	0.07997	0.08019	0.08025	0.08026
6	0.04935	0.07062	0.07493	0.07597	0.07623	0.07630	0.07632
7	0.04367	0.06431	0.06990	0.07125	0.07159	0.07167	0.07170
8	0.03924	0.05828	0.06478	0.06639	0.06679	0.06690	0.06692
9	0.03441	0.05436	0.06057	0.06223	0.06265	0.06276	0.06279

An important question here is the relationship between these numbers and the right-hand side of the estimates in (3.10) and (3.13). To answer this question, the eigenvalues of the operator Q_S have to be determined. It follows from the divergence theorem that

$$\int_\Omega (\mathbf{b}\cdot\nabla u)\overline{v} = -\int_\Omega u(\mathbf{b}\cdot\nabla\overline{v}) - \int_\Omega (\operatorname{div}\mathbf{b})u\overline{v} + \int_{\Gamma_N} (\mathbf{b}\cdot\nu)u\overline{v}\,\mathrm{d}\sigma \quad \left(u, v \in H^1_D(\Omega)\right), \tag{3.15}$$

but in our case $\operatorname{div}\mathbf{b} = 0$ and $0 = (1,0)\cdot(0,\pm 1) = \mathbf{b}\cdot\nu$ on Γ_N. Considering this equation and the definition of the operator $Q_S : H^1_D(\Omega) \to H^1_D(\Omega)$ in equation (3.9),

Q_S can be written as

$$\langle Q_S u, v\rangle_{H_D^1} = \int_\Omega (\mathbf{b}\cdot\nabla u)\overline{v} \qquad \forall\ u, v \in H_D^1(\Omega). \tag{3.16}$$

The eigenvalue problem for Q_S can be formulated in the following way:

$$\left.\begin{aligned} Q_S u &= \lambda u \\ u\big|_{\Gamma_D} &= 0 \end{aligned}\right\} \iff \begin{cases} \langle Q_S u, v\rangle_{H_D^1} = \lambda\,\langle u, v\rangle_{H_D^1} & \forall v \in H_D^1(\Omega) \\ u\big|_{\Gamma_D} = 0. \end{cases} \tag{3.17}$$

Transforming the first equation on the right-hand side and replacing λ by $1/\mu$ we get

$$0 = \int_\Omega \left(-\Delta u - \mu\,(\mathbf{b}\cdot\nabla u) + cu\right)\overline{v} + \int_{\Gamma_N} \frac{\partial u}{\partial \nu}\overline{v} \qquad \forall\ v \in H_D^1(\Omega). \tag{3.18}$$

We have $H_0^1(\Omega) \subset H_D^1(\Omega)$, hence the eigenvalue problem has the form in the cases (a) and (b)

$$\left.\begin{aligned} -\Delta u - \mu\frac{\partial u}{\partial x} + cu &= 0 \\ u\big|_{\partial\Omega} &= 0; \end{aligned}\right\} \tag{3.19}$$

and

$$\left.\begin{aligned} -\Delta u - \mu\frac{\partial u}{\partial x} + cu &= 0 \\ u\big|_{\Gamma_D} = 0, \quad \frac{\partial u}{\partial \nu}\big|_{\Gamma_N} &= 0, \end{aligned}\right\} \tag{3.20}$$

respectively. Let us consider the second problem for instance. We have to find a nonzero function u and some number μ which satisfy equation (3.20) and the two additional boundary conditions. Following the way of calculation for a similar problem in [45, Sec. 2], let us consider an auxiliary equation instead of solving our problem directly:

$$-\Delta v - \mu\frac{\partial v}{\partial x} + cv = \delta(\mu) v \tag{3.21}$$

with the same boundary conditions as in equation (3.20). The eigenfunctions of this problem are also the eigenfunctions for the original problem (3.20) and the values of μ are computable by solving the equation $\delta(\mu) = 0$. It is easy to verify that the functions

$$v_{jk}(x, y) = \exp\left(-\frac{\mu}{2}x\right)\sin(j\pi x)\cos(k\pi y) \qquad \left(j \in \mathbb{N}^+,\ k \in \mathbb{N}\right)$$

are non-zero and satisfy the boundary conditions and equation (3.21) as well with the corresponding numbers

$$\delta_{jk}(\mu) = \left(j^2 + k^2\right)\pi^2 + \frac{\mu^2}{4} + c.$$

The other problem (3.19) with respect to the eigenvalue problem for test problem (a) can be solved in the same way, the functions and δ's are

$$v_{jk}(x,y) = \exp\left(-\frac{\mu}{2}x\right)\sin(j\pi x)\sin(k\pi y) \qquad \left(j,k \in \mathbb{N}^+\right)$$

and

$$\delta_{jk}(\mu) = \left(j^2+k^2\right)\pi^2 + \frac{\mu^2}{4} + c$$

which are formally the same as the previous ones, but the indices are different. Solving equation $\delta(\mu) = 0$ and replacing μ by $1/\lambda$, the eigenvalues of Q_S are

$$\lambda_{jk} = \pm\frac{\mathrm{i}}{2\sqrt{(k^2+j^2)\pi^2 + c}}$$

where $j,k \neq 0$ for problem (a) and $j \neq 0$ for problem (b). Note that the eigenvalues are purely imaginary and accumulate in the origin. Now we can compare the values of Q_k and the upper bound provided by the estimate in Corollary 3.4.

Tab. 3.3: Comparison between the values of Q_k and estimate (3.13)

	problem (a)			problem (b)			$\frac{2}{k}\sum_{i=1}^{k}\lvert\lambda_i(Q_S)\rvert$	
Itr.	64	128	256	64	128	256	(a)	(b)
1	0.0791	0.0792	0.0792	0.1031	0.1031	0.1031	0.2196	0.3033
2	0.0703	0.0704	0.0704	0.0941	0.0941	0.0941	0.2196	0.3033
3	0.0644	0.0645	0.0645	0.0886	0.0886	0.0886	0.1934	0.2754
4	0.0594	0.0595	0.0595	0.0840	0.0840	0.0840	0.1803	0.2615
5	0.0549	0.0550	0.0550	0.0802	0.0803	0.0803	0.1724	0.2406
6	0.0508	0.0509	0.0510	0.0762	0.0763	0.0763	0.1671	0.2267
7	0.0473	0.0474	0.0474	0.0716	0.0717	0.0717	0.1592	0.2144
8	0.0441	0.0443	0.0443	0.0668	0.0669	0.0669	0.1533	0.2053
9	0.0414	0.0417	0.0438	0.0627	0.0628	0.0628	0.1474	0.1981

Table 3.3 shows that the computational results are approximately three times better than the predicted theoretical estimate in both cases.

Experiment 2 ($c_S \neq 1 = c_L$) Turning one's attention to preconditioning with not the symmetric part of L, i.e. $c_S \neq c_L$, surprisingly similar results are shown. In this case the preconditioner is different from the one the theorems are about. The surprise is that nearly the same convergence results are shown with using the GCG-LS(0) algorithm (which now does not coincide with the full version), although the conditions for the convergence theorems are not satisfied.

Tab. 3.4: Values of Q_k boundary conditions (b), $c_S \neq 1 = c_L$.

	1/h=32				1/h=128			
Itr.	$c_S = 0$	$c_S = 0.5$	$c_S = 1.5$	$c_S = 5$	$c_S = 0$	$c_S = 0.5$	$c_S = 1.5$	$c_S = 5$
1	0.1032	0.1023	0.1047	0.1331	0.1034	0.1025	0.1049	0.1332
2	0.0940	0.0939	0.0939	0.1160	0.0943	0.0942	0.0941	0.1162
3	0.0924	0.0897	0.0888	0.1005	0.0926	0.0900	0.0890	0.1008
4	0.0911	0.0856	0.0847	0.0923	0.0914	0.0858	0.0849	0.0923
5	0.0897	0.0845	0.0833	0.0988	0.0899	0.0847	0.0835	0.0985
6	0.0937	0.0857	0.0846	0.0989	0.0939	0.0860	0.0848	0.0990
7	0.0945	0.0869	0.0860	0.0967	0.0947	0.0871	0.0862	0.0968
8	0.0926	0.0879	0.0866	0.0902	0.0929	0.0881	0.0868	0.0906
9	0.0945	0.0872	0.0865	0.0896	0.0948	0.0875	0.0868	0.0896

The results show that the superlinear convergence is not realized during 8-9 iterations, but the numbers are very close to that rate, even when c_S is large. Let us solve problem (b) numerically and set $c_L = 1$. The case $c_S = 1$ has been investigated already. Table 3.4 shows the results of numerical computations for several other constants c_S. For a fixed value of c_S one can also see the mesh independence by comparing the numbers in the appropriate columns.

Experiment 3 ($c_S \neq 0 = c_L$) The same result is shown in Table 3.5, when the roles of c has been transposed, i.e. $c_L = 0$ and c_S varies. In this case there is no zeroth-order term in the operator L, but this term has been put with some constant c_S in S. The constant c_S can be negative and for this case the results are similar as columns for $c_S = \pm 0.5$ show. When negative c_S is used, then the coercivity condition (*iii*) in Assumptions 3.1 is not satisfied.

Tab. 3.5: Values of Q_k boundary conditions (b), $c_S \neq 0 = c_L$.

	1/h=32				1/h=128			
Itr.	$c_S = -\frac{1}{2}$	$c_S = \frac{1}{2}$	$c_S = 1$	$c_S = 5$	$c_S = -\frac{1}{2}$	$c_S = \frac{1}{2}$	$c_S = 1$	$c_S = 5$
1	0.1072	0.1100	0.1132	0.1551	0.1075	0.1102	0.1134	0.1552
2	0.0986	0.0985	0.0987	0.1383	0.0989	0.0987	0.0990	0.1384
3	0.0945	0.0933	0.0944	0.1224	0.0948	0.0936	0.0946	0.1226
4	0.0900	0.0891	0.0938	0.1286	0.0903	0.0893	0.0940	0.1286
5	0.0890	0.0876	0.0929	0.1393	0.0892	0.0878	0.0931	0.1396
6	0.0908	0.0894	0.0956	0.1302	0.0910	0.0897	0.0958	0.1304
7	0.0919	0.0910	0.0984	0.1226	0.0921	0.0912	0.0987	0.1229
8	0.0925	0.0910	0.0970	0.1256	0.0928	0.0912	0.0972	0.1258
9	0.0919	0.0909	0.0968	0.1275	0.0922	0.0912	0.0971	0.1280

Experiment 4 Not every symmetric operator has the same good property. The pur-

pose of this experiment is to prove the importance of the required boundary conditions of S with respect to the given operator L. Let us consider equation (3.14) with boundary conditions (b). Let S be the symmetric part of L, but with the different boundary conditions (a). The values of q_k in Table 3.6 show that the algorithm does not even converge in this case, as theoretical results for equivalent operators predicted in [46]. The reason is that S and L must have Dirichlet boundary conditions on the same portion of the boundary (cf. [46, 31]) and this is not realized in this case. The norm of the error vector does not converge to zero, this procedure is useless.

Tab. 3.6: Values of q_k, boundary conditions (b) in L, boundary conditions (a) in S

	1/h					
Itr.	4	8	16	32	64	128
1	0.8338	0.8064	0.7989	0.7970	0.7965	0.7964
2	0.8321	0.8038	0.7961	0.7941	0.7936	0.7934
3	0.8321	0.8038	0.7960	0.7940	0.7935	0.7934
4	0.8321	0.8038	0.7960	0.7940	0.7935	0.7934
5	0.8321	0.8038	0.7960	0.7940	0.7935	0.7934

Experiment 5 ($\mathbf{b} = (\eta, 0)$) Finally problems with large convection term are considered. In the previous experiments only 8-10 iterations were needed to reach a prescribed accuracy, say $\|e_9\|_{\mathbf{S}_h} \leq 10^{-13}$. The number of the required iterations for larger η grows. Let us fix $c_L = c_S = 1$ and run the algorithm with convection parameter $\eta = (1), 10, 20, \ldots, 50$.

Tab. 3.7: Values of Q_k, boundary conditions (b), $\eta = 10$

	1/h					
Itr.	4	8	16	32	64	128
1	0.6047	0.6463	0.6562	0.6587	0.6593	0.6594
2	0.5742	0.6202	0.6309	0.6335	0.6342	0.6344
3	0.5316	0.5848	0.5965	0.5994	0.6001	0.6002
14	0.1859	0.3726	0.3978	0.4118	0.4160	0.4171
15	0.0983	0.3676	0.3898	0.3952	0.3993	0.4005
16		0.3612	0.3890	0.3846	0.3864	0.3873
23		0.2923	0.3482	0.3508	0.3531	0.3535
24		0.2840	0.3411	0.3489	0.3437	0.3445
25			0.3361	0.3474	0.3408	0.3472
26			0.3302	0.3421	0.3481	0.3585
27				0.3468	0.3575	0.3591
28				0.3570	0.3598	0.3544

See Table 3.7 for the numerical results when the convection term η is larger. The

required number of iterations rapidly grows, but the superlinear convergence property still holds. If the accuracy is fixed to 10^{-8}, then the number of needed iterations is shown in Table 3.8 for different values of $\mathbf{b} = (\eta, 0)$ and mesh parameters h. If we set aside from the coarse mesh parameters $h^{-1} \leq 8$, the other partitions show similar behavior for large values of η, as it turns out from Table 3.8. Considering the rows for $h^{-1} = 32, 64, 128$ and 256, the number of iterations grows together, i.e. the mesh independence property is also valid. Nevertheless, for problems with large η the required number of iterations is also large and the problem might be handled with proper modifications this algorithm, such as using a mixed formulation or involving coefficients that only vary on boundary layers.

Tab. 3.8: Required number of iterations, $\|e_k\|_{\mathbf{S}_h} \leq 10^{-8}$.

	η							
$1/h$	1	10	20	30	40	50	100	500
8	7	17	24	30	37	40	62	119
16	7	18	27	36	44	51	91	338
32	7	18	29	38	46	55	99	415
64	7	18	29	39	46	55	99	430
128	7	18	29	39	46	55	99	439
256	7	18	29	40	46	55	99	444

Summing up, the conjugate gradient method with symmetric and symmetric part preconditioning has proven an efficient algorithm for convection-diffusion problems with small or medium convection term.

3.2 Equations with nonhomogeneous mixed boundary conditions

In this section the PCG method is applied to solving convection-diffusion equations with nonhomogeneous mixed boundary conditions. Using the approach of equivalent and compact-equivalent operators in Hilbert space, it is shown that for a wide class of elliptic problems the superlinear convergence of the obtained preconditioned CGM is mesh independent under FEM discretization.

The theory of compact-equivalent operators has been summarized in Section 2.3. This was based on [10], where the CGN algorithm was applied and superlinear convergence estimate was obtained for elliptic equations with homogeneous mixed boundary conditions. Here we complete those results for convection-diffusion equations with nonhomogeneous mixed boundary conditions. In this case, the main difficulty arises from the proper definition of the corresponding unbounded operator, since including the boundary conditions in the domain of L results that the domain of the operator does

not form a subspace in H. Hence it should consist of a pair of operators defined on the domain itself and on the Neumann boundary. Here the CGN method will be used instead of the GCG-LS algorithm, since the compact-equivalence property will be used and we can get rid of the restrictive normality condition of Section 2.4.

3.2.1 Coercive elliptic differential operators

Let us consider the elliptic partial differential equation

$$\left.\begin{aligned} -\operatorname{div}(A\,\nabla u) + \mathbf{b}\cdot\nabla u + cu &= g \\ \frac{\partial u}{\partial \nu_A} + \alpha u\big|_{\Gamma_N} &= \gamma \\ u\big|_{\Gamma_D} &= 0, \end{aligned}\right\} \tag{3.22}$$

where

$$\frac{\partial u}{\partial \nu_A} = A\nu\cdot\nabla u$$

is the weighted form of the normal derivative. We assume that the following assumptions are satisfied:

Assumptions 3.5. *Suppose that*

(i) $\Omega \subset \mathbb{R}^d$ *is a bounded piecewise* C^1 *domain;* Γ_D, Γ_N *are disjoint open measurable subparts of* $\partial\Omega$ *such that* $\partial\Omega = \overline{\Gamma}_D \cup \overline{\Gamma}_N$*;*

(ii) $A \in L^\infty(\overline{\Omega}, \mathbb{R}^{d\times d})$ *and for all* $x \in \overline{\Omega}$ *the matrix* $A(x)$ *is symmetric; further,* $\mathbf{b} \in W^{1,\infty}(\Omega)^d$, $c \in L^\infty(\Omega)$, $\alpha \in L^\infty(\Gamma_N)$*;*

(iii) we have the coercivity properties

$$\exists\ p > 0 \text{ such that } A(x)\xi\cdot\xi \geq p\,|\xi|^2 \qquad \forall\ x \in \overline{\Omega},\ \xi \in \mathbb{R}^d \tag{3.23}$$

$$\hat{c} := c - \frac{1}{2}\operatorname{div}\mathbf{b} \geq 0 \text{ in } \Omega, \qquad \hat{\alpha} := \alpha + \frac{1}{2}(\mathbf{b}\cdot\nu) \geq 0 \text{ on } \Gamma_N; \tag{3.24}$$

(iv) either $\Gamma_D \neq \emptyset$*, or* $\hat{c}$ *or* $\hat{\alpha}$ *has a positive lower bound.*

The definition of the operator L which corresponds to equation (3.22) has to be understood as a pair of operators: one acts on Ω and the other one acts on the Neumann boundary. Formally we have

$$L \equiv \begin{pmatrix} M \\ P \end{pmatrix}, \quad L\begin{pmatrix} u \\ \eta \end{pmatrix} = \begin{pmatrix} Mu \\ P\eta \end{pmatrix} = \begin{pmatrix} -\operatorname{div}(A\,\nabla u) + \mathbf{b}\cdot\nabla u + cu \\ \dfrac{\partial \eta}{\partial \nu_A} + \alpha\eta\big|_{\Gamma_N} \end{pmatrix}. \tag{3.25}$$

Let us define a symmetric elliptic operator on the same domain in an analogous way:

$$S \equiv \begin{pmatrix} N \\ R \end{pmatrix}, \quad S \begin{pmatrix} u \\ \eta \end{pmatrix} = \begin{pmatrix} Nu \\ R\eta \end{pmatrix} = \begin{pmatrix} -\operatorname{div}(G\,\nabla u) + \sigma u \\ \dfrac{\partial \eta}{\partial \nu_G} + \beta \eta\big|_{\Gamma_N} \end{pmatrix} \tag{3.26}$$

satisfying similar assumptions as of L:

Assumptions 3.6. *Suppose that*

(i) substituting G for A, Ω, Γ_D, Γ_N and G satisfy Assumptions 3.5;

(ii) $\sigma \in L^\infty(\Omega)$, $\sigma \geq 0$, $\beta \in L^\infty(\Gamma_N)$, $\beta \geq 0$; further, if $\Gamma_D \neq \emptyset$, then σ or β has a positive lower bound.

If $\gamma = 0$ in equation (3.22) then under Assumptions 3.5-3.6 the operator L is S-bounded and S-coercive, which has been proved in [11, Prop. 3.9]. Here we extend the scope of that result to the nonhomogeneous case. Let us consider the differential equation (3.22) again. We are interested in solving the analogous operator equation

$$L \begin{pmatrix} u \\ u\big|_{\Gamma_N} \end{pmatrix} = \begin{pmatrix} g \\ \gamma \end{pmatrix}, \tag{3.27}$$

which is the appropriately modified version of the operator equation (2.12). Now we would like to apply the framework developed in Section 2.3 for the elliptic operator L. The Hilbert space H is defined as the product space

$$H = L^2(\Omega) \times L^2(\Gamma_N)$$

endowed with the inner product

$$\left\langle \begin{pmatrix} u \\ \eta \end{pmatrix}, \begin{pmatrix} v \\ \zeta \end{pmatrix} \right\rangle_H := \langle u, v \rangle_{L^2(\Omega)} + \langle \eta, \zeta \rangle_{L^2(\Gamma_N)}.$$

We define the energy space

$$H_S := \left\{ \begin{pmatrix} u \\ u\big|_{\Gamma_N} \end{pmatrix} : \; u \in H^1(\Omega), \; u\big|_{\Gamma_D} = 0 \right\}$$

with the inner product

$$\begin{aligned}\left\langle \begin{pmatrix} u \\ u|_{\Gamma_N} \end{pmatrix}, \begin{pmatrix} v \\ v|_{\Gamma_N} \end{pmatrix} \right\rangle_S &= \left\langle S \begin{pmatrix} u \\ u|_{\Gamma_N} \end{pmatrix}, \begin{pmatrix} v \\ v|_{\Gamma_N} \end{pmatrix} \right\rangle_H = \left\langle \begin{pmatrix} Nu \\ Ru|_{\Gamma_N} \end{pmatrix}, \begin{pmatrix} v \\ v|_{\Gamma_N} \end{pmatrix} \right\rangle_H \\ &= \langle Nu, v \rangle_{L^2(\Omega)} + \left\langle Ru|_{\Gamma_N}, v|_{\Gamma_N} \right\rangle_{L^2(\Gamma_N)} \\ &= \left[\int_\Omega (G\, \nabla u \cdot \nabla v + \sigma uv) - \int_{\Gamma_N} \frac{\partial u}{\partial \nu_G} v \right] + \int_{\Gamma_N} \left(\frac{\partial u}{\partial \nu_G} + \beta u \right) v \\ &= \int_\Omega (G\, \nabla u \cdot \nabla v + \sigma uv) + \int_{\Gamma_N} \beta uv. \end{aligned} \tag{3.28}$$

Proposition 3.7. *If Assumptions 3.5-3.6 hold, then the operator L is S-bounded and S-coercive in H, i.e. $L \in BC_S\left(L^2(\Omega) \times L^2(\Gamma_N)\right)$.*

Proof. Following [11], we have to verify the properties listed in Definition 2.37. The domain of L is

$$D(L) := \left\{ \begin{pmatrix} u \\ u|_{\Gamma_N} \end{pmatrix} :\ u \in H^2(\Omega),\ u|_{\Gamma_D} = 0 \right\},$$

$D(L) \subset H_S$ and $D(L)$ is dense in H_S in the S-inner product. Since the trace of an H^2-function on the Neumann boundary belongs to $L^2(\Gamma_N)$, we have $L : D(L) \subset H \to H$, i.e. L is well-defined on H. Using Green's formula we have

$$\begin{aligned}\left\langle L \begin{pmatrix} u \\ u|_{\Gamma_N} \end{pmatrix}, \begin{pmatrix} v \\ v|_{\Gamma_N} \end{pmatrix} \right\rangle_H &= \left\langle \begin{pmatrix} Mu \\ Pu|_{\Gamma_N} \end{pmatrix}, \begin{pmatrix} v \\ v|_{\Gamma_N} \end{pmatrix} \right\rangle_H \\ &= \langle Mu, v \rangle_{L^2(\Omega)} + \left\langle Pu|_{\Gamma_N}, v|_{\Gamma_N} \right\rangle_{L^2(\Gamma_N)} \\ &= \int_\Omega (A\, \nabla u \cdot \nabla v + (\mathbf{b} \cdot \nabla u)\, v + cuv) + \int_{\Gamma_N} \alpha uv. \end{aligned} \tag{3.29}$$

Using this, properties 2. and 3. in Definition 2.37 have to be verified, but since formally we have the same expressions for the bilinear form of L and for the S-norm as in the homogeneous case, from here the proof goes exactly the same way as in [11, Prop. 3.9], so we omit the further details. □

It follows from Green's formula that the weak solution of (3.27) described in Definition 2.42 is nothing else than the weak solution of the PDE (3.22) in the usual sense, i.e. for a given pair of functions $g \in L^2(\Omega)$ and $\gamma \in L^2(\Gamma_N)$ we have

$$\left\langle L_S \begin{pmatrix} u \\ u|_{\Gamma_N} \end{pmatrix}, \begin{pmatrix} v \\ v|_{\Gamma_N} \end{pmatrix} \right\rangle_S = \left\langle \begin{pmatrix} g \\ \gamma \end{pmatrix}, \begin{pmatrix} v \\ v|_{\Gamma_N} \end{pmatrix} \right\rangle_H \qquad \left(\begin{pmatrix} v \\ v|_{\Gamma_N} \end{pmatrix} \in H_S \right)$$

if and only if

$$\begin{aligned}\left\langle L_S \begin{pmatrix} u \\ u\big|_{\Gamma_N} \end{pmatrix}, \begin{pmatrix} v \\ v\big|_{\Gamma_N} \end{pmatrix} \right\rangle_S &= \int_\Omega \left(A\ \nabla u \cdot \nabla v + (\mathbf{b} \cdot \nabla u)\, v + cuv\right) + \int_{\Gamma_N} \alpha uv \\ &= \int_\Omega gv + \int_{\Gamma_N} \gamma v = \langle g, v\rangle_{L^2(\Omega)} + \left\langle \gamma, v\big|_{\Gamma_N} \right\rangle_{L^2(\Gamma_N)} \qquad (3.30) \\ &= \left\langle \begin{pmatrix} g \\ \gamma \end{pmatrix}, \begin{pmatrix} v \\ v\big|_{\Gamma_N} \end{pmatrix} \right\rangle_H ,\end{aligned}$$

that is the weak solution is the uniquely existing solution of

$$\begin{aligned}&\int_\Omega \left(A\ \nabla u \cdot \nabla v + (\mathbf{b} \cdot \nabla u)\, v + cuv\right) + \int_{\Gamma_N} \alpha uv \\ &\qquad = \int_\Omega gv + \int_{\Gamma_N} \gamma v \qquad \left(v \in H^1(\Omega),\ v\big|_{\Gamma_D} = 0\right). \qquad (3.31)\end{aligned}$$

Remark 3.8. The energy space H_S can be identified with the space

$$H^1_D(\Omega) = \left\{ u \in H^1(\Omega) :\ u\big|_{\Gamma_D} = 0 \right\},$$

with the obvious correspondence $u \mapsto \left(u, u\big|_{\Gamma_N}\right)$, which is the usual energy space for the homogeneous differential operator.

Consider two elliptic differential operators of the form (3.22) with homogeneous Dirichlet boundary conditions on the same part of the boundary. Then the compact-equivalence of these operators can be characterized as follows, see [10, Prop. 3.1].

Proposition 3.9. *Elliptic differential operators satisfying Assumptions 3.5 are compact-equivalent in $H^1_D(\Omega)$ if and only if their principal parts coincide up to some constant $\mu > 0$.*

3.2.2 Symmetric compact-equivalent preconditioners and mesh independent superlinear convergence

Now we consider the finite element discretization of problem (3.22), where the corresponding operator L is S-bounded and S-coercive, $g \in L^2(\Omega)$, $\gamma \in L^2(\Gamma_N)$. We note that the finite element method fits naturally in the framework developed in Section 2.3, since we are looking for the weak solution described in Definition 2.42, which is nothing else than the variational form (3.31) of equation (3.22).

Let

$$V_h = \operatorname{span}\{\varphi_1, \varphi_2, \ldots, \varphi_n\} \subset H^1_D(\Omega)$$

be a given n-dimensional subspace. The finite element solution $u_h \in V_h$ is $u_h = \sum_{j=1}^{n} c_i \varphi_j$, where $\mathbf{c} = (c_1, c_2, \ldots, c_n) \in \mathbb{R}^n$ is the solution of the linear system

$$\mathbf{L}_h \mathbf{c} = \mathbf{d}_h, \tag{3.32}$$

where

$$(\mathbf{L}_h)_{ij} = \int_\Omega \left(A\ \nabla\varphi_i \cdot \nabla\varphi_j + (\mathbf{b} \cdot \nabla\varphi_j)\,\varphi_i + c\varphi_i\varphi_j\right) + \int_{\Gamma_N} \alpha\varphi_i\varphi_j \tag{3.33}$$

and

$$(\mathbf{d}_h)_j = \int_\Omega g\varphi_j + \int_{\Gamma_N} \gamma\varphi_j. \tag{3.34}$$

Let us take the symmetric operator described in Definition 2.37 and introduce the stiffness matrix of S in H_S

$$(\mathbf{S}_h)_{ij} = \langle \varphi_i, \varphi_j \rangle_S = \int_\Omega \left(G\ \nabla\varphi_i \cdot \nabla\varphi_j + \sigma\varphi_i\varphi_j\right) + \int_{\Gamma_N} \beta\varphi_i\varphi_j.$$

To solve the preconditioned system

$$\mathbf{S}_h^{-1}\mathbf{L}_h\mathbf{c} = \mathbf{S}_h^{-1}\mathbf{d}_h \tag{3.35}$$

one can turn to the conjugate gradient methods in Section 2.2 using the $\mathbf{S}_h$-inner product $\langle \cdot, \cdot \rangle_{\mathbf{S}_h}$.

A sometimes good strategy to solve (3.32) is to choose the preconditioner as the symmetric part of $\mathbf{L}_h$, as it has done in the previous subsection for similar equations with homogeneous boundary conditions. Let us define

$$\mathbf{S}_h := \frac{\mathbf{L}_h + \mathbf{L}_h^T}{2}, \qquad \mathbf{Q}_h := \frac{\mathbf{L}_h - \mathbf{L}_h^T}{2},$$

the symmetric and antisymmetric parts of $\mathbf{L}_h$ and chose the matrix $\mathbf{S}_h$ as preconditioner for $\mathbf{L}_h$. In this case the preconditioned equation (3.35) becomes

$$\left(\mathbf{I}_h + \mathbf{S}_h^{-1}\mathbf{Q}_h\right)\mathbf{c} = \mathbf{S}_h^{-1}\mathbf{d}_h, \tag{3.36}$$

where the matrix $\mathbf{S}_h^{-1}\mathbf{Q}_h$ is antisymmetric in $\langle \cdot, \cdot \rangle_{\mathbf{S}_h}$, thus the GCG-LS algorithm 2.28 coincides with the truncated GCG-LS(0) algorithm 2.29. Now we have to define an appropriate elliptic operator S such that the stiffness matrix $\mathbf{S}_h$ becomes the symmetric

part of $\mathbf{L}_h$, which belongs to the operator L defined in (3.22). Just as in Subsection 3.1 the symmetric part has to be defined in weak sense (cf. Subsection 2.4.2).

For the given elliptic equation (3.22), its symmetric part can be constructed as

$$S \begin{pmatrix} u \\ u|_{\Gamma_N} \end{pmatrix} \equiv \begin{pmatrix} -\operatorname{div}(A\,\nabla u) + \hat{c}u \\ \dfrac{\partial u}{\partial \nu_A} + \hat{\alpha} u|_{\Gamma_N} \end{pmatrix} \tag{3.37}$$

where

$$\hat{c} = c - \frac{1}{2}\operatorname{div}\mathbf{b}, \qquad \hat{\alpha} = \alpha + \frac{1}{2}(\mathbf{b}\cdot\nu)\,. \tag{3.38}$$

Since L satisfies Assumptions 3.5, it is easy to see that S satisfies Assumptions 3.6. The corresponding S-inner product on H_S is

$$\left\langle \begin{pmatrix} u \\ u|_{\Gamma_N} \end{pmatrix}, \begin{pmatrix} v \\ v|_{\Gamma_N} \end{pmatrix} \right\rangle_S = \int_\Omega (A\,\nabla u\cdot\nabla v + \hat{c}uv) + \int_{\Gamma_N} \hat{\alpha} uv. \tag{3.39}$$

Using the divergence theorem and Green's formula, it is easy to check that

$$\begin{aligned} &\left\langle \begin{pmatrix} u \\ u|_{\Gamma_N} \end{pmatrix}, \begin{pmatrix} v \\ v|_{\Gamma_N} \end{pmatrix} \right\rangle_S \\ &\qquad = \frac{1}{2}\left[\left\langle L_S \begin{pmatrix} u \\ u|_{\Gamma_N} \end{pmatrix}, \begin{pmatrix} v \\ v|_{\Gamma_N} \end{pmatrix} \right\rangle_S + \left\langle \begin{pmatrix} u \\ u|_{\Gamma_N} \end{pmatrix}, L_S \begin{pmatrix} v \\ v|_{\Gamma_N} \end{pmatrix} \right\rangle_S \right], \end{aligned} \tag{3.40}$$

that is the corresponding matrix $\mathbf{S}_h$ is indeed the symmetric part of $\mathbf{L}_h$, hence the operator L_S can be decomposed as

$$L_S = I + Q_S,$$

where I is the identity and Q_S is an antisymmetric operator on H_S defined by

$$\begin{aligned} &\left\langle Q_S \begin{pmatrix} u \\ u|_{\Gamma_N} \end{pmatrix}, \begin{pmatrix} v \\ v|_{\Gamma_N} \end{pmatrix} \right\rangle_S = \\ &\qquad = \frac{1}{2}\left[\left\langle L_S \begin{pmatrix} u \\ u|_{\Gamma_N} \end{pmatrix}, \begin{pmatrix} v \\ v|_{\Gamma_N} \end{pmatrix} \right\rangle_S - \left\langle \begin{pmatrix} u \\ u|_{\Gamma_N} \end{pmatrix}, L_S \begin{pmatrix} v \\ v|_{\Gamma_N} \end{pmatrix} \right\rangle_S \right] \\ &\qquad = \frac{1}{2}\int_\Omega \big((\mathbf{b}\cdot\nabla u)v - u(\mathbf{b}\cdot\nabla v)\big). \end{aligned} \tag{3.41}$$

Consider again the differential equation (3.22) with the corresponding operator L in (3.25) and preconditioner S in (3.26) and assume that $A = G$, then it follows from Proposition 3.9 that L and S are compact-equivalent with $\mu = 1$, thus (2.19) holds.

Now let us consider the preconditioned equation (3.36), when $\mathbf{L}_h$ and $\mathbf{S}_h$ now come from the elliptic operators L and S, $\mathbf{Q}_h = \mathbf{L}_h - \mathbf{S}_h$ and now S is not necessarily the symmetric part of L, i.e. S has general coefficients as in (3.26). In this case the operator Q_S is defined as

$$\left\langle Q_S \begin{pmatrix} u \\ u|_{\Gamma_N} \end{pmatrix}, \begin{pmatrix} v \\ v|_{\Gamma_N} \end{pmatrix} \right\rangle_S = \left\langle L_S \begin{pmatrix} u \\ u|_{\Gamma_N} \end{pmatrix}, \begin{pmatrix} v \\ v|_{\Gamma_N} \end{pmatrix} \right\rangle_S - \left\langle \begin{pmatrix} u \\ u|_{\Gamma_N} \end{pmatrix}, \begin{pmatrix} v \\ v|_{\Gamma_N} \end{pmatrix} \right\rangle_S$$
$$= \int_\Omega \big((\mathbf{b} \cdot \nabla u)v + (c - \sigma)uv \big) + \int_{\Gamma_N} (\alpha - \beta)uv, \quad (3.42)$$

which coincides with (3.41) if $\sigma = \hat{c}$ and $\beta = \hat{\alpha}$, where these coefficients are given in (3.38).

When symmetric part preconditioning is used as in Subsection 2.4.2, i.e. when the preconditioner S is defined as in (3.37), then the antisymmetric part $Q_S \in B(H_S)$ – which is given in (3.41) – is compact normal operator and the matrix $\mathbf{S}_h^{-1}\mathbf{Q}_h$ is $\mathbf{S}_h$-normal with respect to $\langle \cdot, \cdot \rangle_{\mathbf{S}_h}$. In this case the superlinear convergence estimate (3.13) holds, and the GCG-LS method reduces to the truncated GCG-LS(0) algorithm 2.29.

When S is not the symmetric part of L, then S is given as in (3.26) and $Q_S \in B(H_S)$ is defined as (3.42). Now the conditions of Theorem 2.44 are satisfied, thus the CGN algorithm 2.33 provides a similar mesh independent superlinear convergence result (with appropriately modified Q_S).

Corollary 3.10. *With Assumptions 3.5 and 3.6 and $A = G$, the CGN algorithm 2.33 for system* (3.36) *yields*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}} \right)^{1/k} \leq \frac{2}{m^2} \left(\frac{1}{k} \sum_{i=1}^{k} \big(|\lambda_i(Q_S^* + Q_S)| + \lambda_i(Q_S^* Q_S) \big) \right) \xrightarrow{k \to \infty} 0,$$

where $m > 0$ comes from the S-coercivity of L in Proposition 3.7.

3.2.3 Numerical experiments

We would like to illustrate the obtained mesh independent superlinear convergence results with a simple numerical example using symmetric part preconditioning. The test problem is the following elliptic convection-diffusion equation

$$\left.\begin{aligned} -\Delta u + \frac{\partial u}{\partial x} + cu &= g \\ u\big|_{\Gamma_D} = 0, \quad \frac{\partial u}{\partial \nu} + \alpha u\big|_{\Gamma_N} &= \gamma. \end{aligned}\right\} \quad (3.43)$$

The parameters of this special model problem has the following properties:

(i) $\Omega = [0,1] \times [0,1]$ is the unit square. The homogeneous Dirichlet boundary condition is given on $\Gamma_D = \{(x,y) \in \partial\Omega : \ x = 0 \text{ or } x = 1\}$;

(ii) $\mathbf{b} = (1,0)$, $c = 1$ and $\alpha = 1$ are constants;

(iii) g and γ are polynomials.

One can verify that Assumptions 3.5 for the general problem are satisfied. Using (3.37) the coefficients $\hat{c}$, $\hat{\alpha}$ in operator S can be readily calculated. Owing to the symmetric part preconditioning strategy, the truncated GCG-LS(0) algorithm 2.29 can be used instead of the full algorithm. The superlinear convergence of the algorithm is provided by the compact-equivalence of L and S with $\mu = 1$. Since we have the decomposition (2.19) with a compact antisymmetric operator Q_S, the truncated algorithm yields the mesh independent convergence estimate

$$\left(\frac{\|e_k\|_{\mathbf{S}_h}}{\|e_0\|_{\mathbf{S}_h}} \right)^{1/k} \leq \frac{2}{k} \sum_{j=1}^{k} |\lambda_j(Q_S)| \xrightarrow{k\to\infty} 0, \tag{3.44}$$

since $m = 1$ and the $\mathbf{L}_h$-norm equals to the $\mathbf{S}_h$-norm, as in (3.13).

Remark 3.11. Now Algorithm 2.29 is applied for a system with $A = \mathbf{S}_h^{-1}\mathbf{L}_h$, thus r_0 is the solution of equation $\mathbf{S}_h r_0 = \mathbf{L}_h u_0 - b$, similarly the calculation of the vector $z_k := Ad_k$ inside the loop leads to the solution of the auxiliary problem $\mathbf{S}_h z_k = \mathbf{L}_h d_k$. Considering the meaning of the matrices $\mathbf{S}_h$ and $\mathbf{L}_h$, the vectors r_0 and d_k are the finite element solution of the problems

$$\left\{ \begin{array}{rcl} -\Delta r_0 + \hat{c} r_0 & = & -\Delta u_0 + \mathbf{b} \cdot \nabla u_0 + c u_0 - g \\ \dfrac{\partial r_0}{\partial \nu} + \hat{\alpha} r_0\big|_{\Gamma_N} & = & \dfrac{\partial u_0}{\partial \nu} + \alpha u_0\big|_{\Gamma_N} - \gamma \\ r_0\big|_{\Gamma_D} & = & 0 \end{array} \right. \tag{3.45}$$

and

$$\left\{ \begin{array}{rcl} -\Delta z_k + \hat{c} z_k & = & -\Delta d_k + \mathbf{b} \cdot \nabla d_k + c d_k \\ \dfrac{\partial z_k}{\partial \nu} + \hat{\alpha} z_k\big|_{\Gamma_N} & = & \dfrac{\partial d_k}{\partial \nu} + \alpha d_k\big|_{\Gamma_N} \\ z_k\big|_{\Gamma_D} & = & 0, \end{array} \right. \tag{3.46}$$

respectively.

In the numerical experiment piecewise linear elements were used, the stopping criterion was $\|e_k\|_{\mathbf{S}_h} \leq 10^{-13}$. In Table 3.9 Q_k denotes the quotient of the error vectors

according to the left-hand side of estimate (3.44):

$$Q_k := \left(\frac{\|e_k\|_{\mathbf{S}_h}}{\|e_0\|_{\mathbf{S}_h}} \right)^{1/k}.$$

As expected, the numbers in Table 3.9 shows that the convergence is superlinear, i.e. the sequence Q_k tends to zero for any value of the mesh parameter.

Tab. 3.9: Values of Q_k for equation (3.43).

	1/h					
Itr.	8	16	32	64	128	256
1	0.0685	0.0706	0.0711	0.0713	0.0713	0.0713
2	0.0761	0.0786	0.0793	0.0794	0.0795	0.0795
3	0.0724	0.0752	0.0759	0.0760	0.0761	0.0761
4	0.0707	0.0738	0.0746	0.0748	0.0748	0.0748
5	0.0667	0.0698	0.0706	0.0708	0.0709	0.0709
6	0.0634	0.0670	0.0679	0.0682	0.0682	0.0682
7	0.0585	0.0630	0.0641	0.0644	0.0644	0.0644
8	0.0543	0.0597	0.0610	0.0613	0.0614	0.0614
9	0.0508	0.0562	0.0577	0.0580	0.0581	0.0582
10	0.0490	0.0542	0.0556	0.0560	0.0561	0.0561
11			0.0544	0.0551	0.0565	0.0590

The numbers in each row show the boundedness of Q_k as the parameter h increases, which yields the desired mesh independent convergence property. Thus using compact-equivalent preconditioner, the superlinear convergence rate of Algorithm 2.29 is also valid for problems with nonhomogeneous mixed boundary conditions.

When the convection term $\mathbf{b} = (b_1, b_2)$ is large, then the mesh independent superlinear convergence property still holds, although the number of required iterations to reach the prescribed tolerance level increases rapidly. Table 3.10 shows these results for $\mathbf{b} = (\eta, 0)$.

Tab. 3.10: Required number of iterations, $\|e_k\|_{\mathbf{S}_h} \leq 10^{-8}$.

	η							
$1/h$	1	10	20	30	40	50	100	500
8	7	17	23	30	37	40	63	127
16	7	18	27	36	44	53	92	367
32	7	18	28	37	47	56	101	438
64	7	18	28	37	47	57	103	460
128	7	18	29	37	47	58	105	468
256	7	18	30	39	49	58	107	475

The results are very similar to Table 3.8, which shows the iteration numbers for the homogeneous case. Altogether symmetric part preconditioning provides a good approximation of L for mildly convection-dominated problems, further comments on singularly perturbed problems can be found in [10, Sec. 5] and [11, Sec. 9].

3.3 Finite difference approximation for equations with Dirichlet boundary conditions

In this section the goal is to study the same problem in the case of finite difference discretizations, i.e. to study the superlinear convergence of the preconditioned CG iteration under equivalent operator preconditioning and to find mesh independent behaviour. Here an important difference arises between FEM and FDM discretizations, pointed out already in [21]. Namely, the FDM lacks the organized Hilbert space background that FEM is based on, hence a case-by-case study of convergence is required for FDM discretizations with equivalent operator preconditioning. For linear convergence such a work has been started already in [18, 24] and extended in [21, 45, 46].

The present section aims to take a first step to verify mesh independence of superlinear convergence, and hence a model problem with Dirichlet boundary conditions is considered on a simple domain with a uniform FD grid. The required mesh independent bound is proved for a certain class of coefficients, and numerical calculations show similar behaviour for other coefficients as well.

3.3.1 Equivalent operator preconditioning

Let us consider an elliptic convection-diffusion equation

$$\left.\begin{aligned} Lu \equiv -\Delta u + \mathbf{b}\cdot\nabla u + cu &= g \\ u|_{\Gamma_D} &= 0 \end{aligned}\right\} \tag{3.47}$$

on a bounded domain $\Omega \subset \mathbb{R}^d$. We assume that $\mathbf{b} \in C^1(\overline{\Omega})^d$ and $c \in L^\infty(\Omega)$; further, there holds the usual coercivity condition

$$c - \frac{1}{2}\operatorname{div}\mathbf{b} \geq 0. \tag{3.48}$$

Here we focus on regularly perturbed problems. The coercivity condition (3.48) implies that for all $g \in L^2(\Omega)$ problem (3.47) has a unique weak solution in $H_0^1(\Omega)$.

The FDM discretization of (3.47) on a given grid ω_h leads to a linear algebraic system

$$L_h u_h = g_h \tag{3.49}$$

of order N for some $N \in \mathbb{N}$. Our goal is to solve (3.49) by iteration, applying a suitable preconditioned conjugate gradient method. The proposed preconditioner is obtained via a symmetric preconditioning operator

$$Su := -\Delta u + \sigma u \quad \text{for} \quad u\big|_{\partial\Omega} = 0, \tag{3.50}$$

where $\sigma \in L^\infty(\Omega)$, $\sigma \geq 0$: namely, the matrix S_h is defined as the FDM discretization of the operator S on the same grid ω_h. The preconditioned form of the discretized system is

$$S_h^{-1} L_h u_h = f_h \equiv S_h^{-1} g_h. \tag{3.51}$$

Here we are interested in the superlinear convergence property of the PCG Algorithm 2.51, where the operators L, S are replaced by the matrices L_h, S_h, respectively. Denoting by u_h^* the unique solution of (3.49), we study the error vector $e_k = u_k - u_h^*$ and use the norm $\|v_h\|_{L_h}^2 = \operatorname{Re}\langle L_h v_h, v_h\rangle$. The related results are formulated by considering the preconditioned matrix as a perturbation of the identity (see (2.7)). Let us decompose our operators as

$$L = S + Q,$$

that is, letting $\gamma = c - \sigma$,

$$Qu = \mathbf{b} \cdot \nabla u + \gamma u. \tag{3.52}$$

Further, let the matrix Q_h be defined as the FDM discretization of the operator Q on the same grid ω_h as for L in (3.49). Then $L_h = S_h + Q_h$, hence (3.51) can be written as

$$\left(I_h + S_h^{-1} Q_h\right) u_h = f_h \equiv S_h^{-1} g_h, \tag{3.53}$$

where I_h is the corresponding identity matrix. Let us define

$$\varrho := \frac{1}{\left\|L_h^{-1} S_h\right\|_{S_h}} \geq \min_{v_h \neq 0} \frac{\operatorname{Re}\langle L_h v_h, v_h\rangle}{\langle S_h v_h, v_h\rangle}.$$

Then the following convergence result holds (cf. Proposition 2.32, and Theorem 2.53 for the analogous infinite dimensional case):

Theorem 3.12. *The GCG-LS method applied for equation* (3.51) *yields*

$$\left(\frac{\|e_k\|_{L_h}}{\|e_0\|_{L_h}}\right)^{1/k} \leq \frac{2}{\varrho}\left(\frac{1}{k}\sum_{i=1}^{k} \left|\lambda_i(S_h^{-1} Q_h)\right|\right) \qquad (k = 1, 2, \ldots, N), \tag{3.54}$$

where $\lambda_i(S_h^{-1} Q_h)$ *are the ordered eigenvalues of the matrix* $S_h^{-1} Q_h$.

This shows superlinear convergence if the eigenvalues $\lambda_i(S_h^{-1} Q_h)$ accumulate in

zero. When symmetric part preconditioning is used, i.e. $S_h = \frac{1}{2}(L_h + L_h^T)$, then the GCG-LS(0) algorithm is applicable. Further, due to the obvious identity $\langle L_h v_h, v_h\rangle = \langle S_h v_h, v_h\rangle$, we have $\varrho = 1$ in this case in (3.54).

In the case of FEM discretization, when L_h and Q_h are the stiffness matrices of L and Q in a FEM subspace, the analogue of the sequence (3.54) can be estimated in a mesh uniform superlinear way (cf. Corollary 3.4). We demonstrate an analogous result for certain finite difference discretizations. For this we need to find a sequence (ε_k), where $\varepsilon_k \to 0$ independently of h such that for all $h > 0$ the eigenvalues $\lambda_i(S_h^{-1}Q_h)$ satisfy

$$\frac{1}{k}\sum_{i=1}^{k} \left|\lambda_i(S_h^{-1}Q_h)\right| \le \varepsilon_k. \tag{3.55}$$

3.3.2 A model problem and the properties of the eigenvalues

Let us consider a special case of (3.47) which has been analysed in [45] in the context of linear convergence. The convection-diffusion problem

$$\left.\begin{aligned} Lu \equiv -\Delta u + \mathbf{b}\cdot\nabla u + cu &= g \\ u\big|_{\Gamma_D} &= 0 \end{aligned}\right\} \tag{3.56}$$

is posed on the unit square $\Omega := [0,1]^2 \subset \mathbb{R}^2$ with constant coefficients $\mathbf{b} = (b_1, b_2) \in \mathbb{R}^2$ and $c \in \mathbb{R}$. We assume $c \ge 0$, then the coercivity condition (3.48) holds. Similarly, in the preconditioning operator

$$Su := -\Delta u + \sigma u \quad \text{for} \quad u\big|_{\partial\Omega} = 0, \tag{3.57}$$

we set $\sigma \in \mathbb{R}$, $\sigma \ge 0$.

Let ω_h be a uniform grid on $[0,1]^2$, $b_1, b_2 \ge 0$ and let us define upwind or centered differencing for the first order and centered differencing for the second order derivatives, respectively. The upwind scheme now coincides with the backward differencing due to the sign conditions on $\mathbf{b} = (b_1, b_2)$. Denote by n the number of interior gridpoints in each direction, and by $h = 1/(n+1)$ the grid parameter. Let L_h, S_h and Q_h denote the $n^2 \times n^2$ matrices corresponding to the discretizations of L, S and $Q = L - S$, respectively. Then by [45], the eigenvalues

$$\mu_{jm} := \lambda_{jm}(S_h^{-1}Q_h) \tag{3.58}$$

of the preconditioned matrix $S_h^{-1}Q_h$ satisfy

$$-\left(4+\sigma h^2\right)\mu_{jm}+\left((c-\sigma)\,h^2+(b_1+b_2)h\right)$$
$$=-2\left(\left(\mu_{jm}^2-\mu_{jm}b_1h\right)^{1/2}\cos m\pi h+\left(\mu_{jm}^2-\mu_{jm}b_2h\right)^{1/2}\cos j\pi h\right) \quad (3.59)$$

(for $j,m=1,\ldots,n$) in the case of the backward differencing and

$$-\left(4+\sigma h^2\right)\mu_{jm}+(c-\sigma)\,h^2$$
$$=-2\left(\left(\mu_{jm}^2-b_1^2h^2/4\right)^{1/2}\cos m\pi h+\left(\mu_{jm}^2-b_2^2h^2/4\right)^{1/2}\cos j\pi h\right) \quad (3.60)$$

(for $j,m=1,\ldots,n$) in the case of the centered differencing approximation of the first order derivative.

3.3.3 Some mesh independent superlinear convergence results

Since the eigenvalues (3.58) are given with double indexing, in view of (3.55) we wish to find a mesh independent sequence $\varepsilon_k\to 0$ independently of h such that for all $h>0$

$$\frac{1}{k^2}\sum_{j,m=1}^{k}|\mu_{jm}|\le\varepsilon_k.$$

In general, the relations (3.59)-(3.60) lead to fourth order algebraic equations whose roots cannot be handled in explicit form. In what follows, first a special class of coefficients is considered where μ_{jm} are obtained directly and an explicit expression can be derived for ε_k. Then some numerical calculations are given which show favourable convergence rates also for other types of coefficients.

Proposition 3.13. *Let us consider problem* (3.56) *with a convection term* $\mathbf{b}=(b,b)$, *where* $b\in\mathbb{R}^+$ *is arbitrary, and let* $\sigma:=c$ *in* (3.57), *i.e.* S *is the symmetric part of* L. *Then, using either centered or backward differencing, the eigenvalues* $\mu_{jm}:=\lambda_{jm}(S_h^{-1}Q_h)$ *satisfy*

$$\frac{1}{k^2}\sum_{j,m=1}^{k}|\mu_{jm}|\le\varepsilon_k\qquad(k=1,2,\ldots,n), \quad (3.61)$$

where

$$\varepsilon_k:=\frac{2\sqrt{2}b}{k^2}\sum_{j,m=1}^{\lfloor\frac{k+1}{2}\rfloor}\frac{1}{\sqrt{\sigma+4m^2+4j^2}}\xrightarrow{k\to\infty}0 \quad (3.62)$$

and ε_k *is independent of* h.

Proof. In the present case relation (3.60) turns into

$$-\left(4+\sigma h^2\right)\mu_{jm} = -2\left(\mu_{jm}^2 - b^2h^2/4\right)^{1/2}\left(\cos m\pi h + \cos j\pi h\right) \tag{3.63}$$

(for $j, m = 1, \ldots, n$), whose roots are purely imaginary and satisfy

$$|\mu_{jm}| = \frac{bh\,|\cos m\pi h + \cos j\pi h|}{\sqrt{\left(4+\sigma h^2\right)^2 - 4\left(\cos m\pi h + \cos j\pi h\right)^2}}. \tag{3.64}$$

The numerator is at most $2bh$, and in the denominator we can use the estimates

$$\left(4+\sigma h^2\right)^2 \geq 16 + 8\sigma h^2,$$
$$\left(\cos m\pi h + \cos j\pi h\right)^2 \leq 2\left(\cos^2 m\pi h + \cos^2 j\pi h\right) = 4 - 2\left(\sin^2 m\pi h + \sin^2 j\pi h\right).$$

Hence we obtain

$$|\mu_{jm}| \leq \frac{bh}{\sqrt{2\left(\sigma h^2 + \sin^2 m\pi h + \sin^2 j\pi h\right)}}. \tag{3.65}$$

If $1 \leq j, m \leq \frac{k+1}{2} \leq \frac{n+1}{2} = \frac{1}{2h}$, then we can use the estimate $\sin t \geq (2/\pi)t$, whence the expression under the root becomes $2h^2\left(\sigma + 4m^2 + 4j^2\right)$ and we obtain

$$|\mu_{jm}| \leq \frac{b}{\sqrt{2\left(\sigma + 4m^2 + 4j^2\right)}} =: \beta_{jm}. \tag{3.66}$$

If j or m is greater than $\frac{k+1}{2}$ and $k \leq \left\lfloor \frac{n+1}{2} \right\rfloor$, then estimate (3.66) is still valid and $|\mu_{jm}| \leq \beta_{jm}$ holds. Further, there exists an injective mapping $(j, m) \mapsto (j', m')$ from the set of index pairs $I_{12} := \{(j, m) :\ 1 \leq j \leq \frac{k+1}{2},\ \frac{k+1}{2} < m \leq k\}$ to the set $I_{11} := \{(j', m') :\ 1 \leq m', j' \leq \frac{k+1}{2}\}$ such that $\beta_{jm} \leq \beta_{j'm'}$ (such a mapping is $(j, m) \mapsto (j, m - \left\lfloor \frac{k+1}{2} \right\rfloor)$). One can readily check the two other cases for the sets of indices I_{21} and I_{22}, hence estimate

$$\frac{1}{k^2}\sum_{j,m=1}^{k} |\mu_{jm}| \leq \frac{4}{k^2}\sum_{j,m=1}^{\left\lfloor \frac{k+1}{2} \right\rfloor} \beta_{jm}$$

holds, which, together with (3.66), implies the required estimate. Similar argument can be used if $\left\lfloor \frac{n+1}{2} \right\rfloor < k \leq n$. For an arbitrary pair of indices (j, m) from one of the index sets I_{12}, I_{21} or I_{22}, there is a corresponding index pair $(j', m') \in I_{11}$ such that $|\mu_{jm}|$ can be estimated above by $\beta_{j'm'}$. Since the right-hand side of (3.66) tends to 0, and ε_k is constant times the arithmetic mean of this sequence, therefore $\varepsilon_k \to 0$ as well. Finally, ε_k is obviously independent of h.

The case of backward differencing is similar. Relation (3.59) becomes

$$-\left(4+\sigma h^2\right)\mu_{jm}+2bh=-2\left(\mu_{jm}^2-\mu_{jm}bh\right)^{1/2}\left(\cos m\pi h+\cos j\pi h\right), \tag{3.67}$$

(for $j,m=1,\ldots,n$), whose roots are

$$\mu_{jm}=\frac{2bh\left(\left(4+\sigma h^2\right)-\left(c_m+c_j\right)^2\pm \mathrm{i}\left|c_m+c_j\right|\sqrt{2\left(\sigma h^2+2\right)-\left(c_m+c_j\right)^2}\right)}{\left(4+\sigma h^2\right)^2-4\left(c_m+c_j\right)^2},$$

where $c_m=\cos m\pi h$ and $c_j=\cos j\pi h$. By elementary calculation, we have

$$\begin{aligned}\left|\mu_{jm}\right|&=\frac{2bh}{\left(4+\sigma h^2\right)^2-4\left(c_m+c_j\right)^2}\Big[\left(\left(4+\sigma h^2\right)-\left(c_m+c_j\right)^2\right)^2\\&\quad+\left(c_m+c_j\right)^2\left(2(\sigma h^2+2)-\left(c_m+c_j\right)^2\right)\Big]^{1/2}=\frac{2bh}{\sqrt{\left(4+\sigma h^2\right)^2-4(c_m+c_j)^2}},\end{aligned} \tag{3.68}$$

which coincides with the trivial upper bound of (3.64), thus the proof goes on the same way as in the centered differencing case. □

Remark 3.14. The eigenvalue bound (3.66) is almost the same as the one obtained in [9, Subsec. 3.4] for the FEM case, differing only in the constants. Hence we have the same rate as proved there if returning to simple indices in ε_k. Namely, let i_s, j_s ($s\in\mathbb{N}^+$) denote the indices of the eigenvalues ordered as $|\mu_{i_1,j_1}|\geq|\mu_{i_2,j_2}|\geq\ldots$ Then there holds

$$\frac{1}{k}\sum_{s=1}^{k}\left|\mu_{i_s,j_s}\right|\leq\frac{C}{\sqrt{k}}\qquad(k=1,2,\ldots)$$

where $C>0$ is independent of k.

The superlinear and the mesh independent behaviour of the arithmetic mean of $\left|\lambda_{jm}(S_h^{-1}Q_h)\right|$ in (3.61) is shown by the columns and rows of the following tables, respectively. In the first part of Table 3.11 centered differencing is considered for a problem in the setting of Proposition 3.13. The last columns show that the behaviour of the eigenvalue mean is almost the same for backward differencing. Table 3.12 shows similar results for $b_1\neq b_2$. We note that [45] suggests $\sigma=\mathcal{O}(b_1^2+b_2^2)$ as an efficient choice in S, which is in correlation with this table in the sense that a smaller σ in Table 3.12 has produced similar numerical results for $b_1=0$ as a greater σ in Table 3.11 for $b_1>0$.

Tab. 3.11: $c = \sigma = 20$, $b_1 = b_2 = 4$.

	centered differencing				backward differencing			
	$1/h$				$1/h$			
Itr.	16	32	64	128	16	32	64	128
1	0.4413	0.4467	0.4482	0.4486	0.4490	0.4488	0.4487	0.4487
2	0.4413	0.4467	0.4482	0.4486	0.4490	0.4488	0.4487	0.4487
3	0.4037	0.4100	0.4117	0.4122	0.4136	0.4127	0.4124	0.4124
4	0.3849	0.3917	0.3935	0.3940	0.3960	0.3946	0.3943	0.3942
5	0.3736	0.3807	0.3826	0.3831	0.3854	0.3838	0.3834	0.3833
6	0.3661	0.3734	0.3753	0.3758	0.3783	0.3766	0.3761	0.3760
7	0.3523	0.3601	0.3622	0.3627	0.3656	0.3636	0.3631	0.3629
15	0.2903	0.3004	0.3031	0.3038	0.3090	0.3053	0.3043	0.3041
16	0.2856	0.2958	0.2986	0.2993	0.3047	0.3009	0.2999	0.2996
17	0.2798	0.2903	0.2931	0.2938	0.2995	0.2955	0.2944	0.2941
63	0.1710	0.1888	0.1936	0.1948	0.2078	0.1984	0.1960	0.1954
64	0.1698	0.1877	0.1924	0.1937	0.2069	0.1974	0.1949	0.1943
65	0.1685	0.1866	0.1914	0.1926	0.2059	0.1963	0.1939	0.1932
255	0.0676	0.1020	0.1109	0.1132	0.1436	0.1210	0.1157	0.1144
256	0.0673	0.1018	0.1107	0.1130	0.1435	0.1208	0.1155	0.1142
257		0.1016	0.1106	0.1128		0.1206	0.1153	0.1140

Tab. 3.12: $c = \sigma = 4$, $b_1 = 0$, $b_2 = 4$.

	centered differencing				backward differencing			
	$1/h$				$1/h$			
Itr.	16	32	64	128	16	32	64	128
1	0.4033	0.4086	0.4100	0.4104	0.4103	0.4104	0.4105	0.4105
2	0.4033	0.4086	0.4100	0.4104	0.4103	0.4104	0.4105	0.4105
3	0.3577	0.3630	0.3644	0.3648	0.3662	0.3653	0.3650	0.3650
4	0.3349	0.3402	0.3417	0.3420	0.3442	0.3427	0.3423	0.3422
5	0.3198	0.3262	0.3279	0.3283	0.3296	0.3288	0.3286	0.3285
6	0.3097	0.3168	0.3187	0.3192	0.3198	0.3195	0.3194	0.3194
7	0.2948	0.3024	0.3044	0.3049	0.3056	0.3053	0.3052	0.3051
15	0.2315	0.2410	0.2435	0.2442	0.2459	0.2448	0.2445	0.2445
16	0.2265	0.2364	0.2391	0.2398	0.2413	0.2403	0.2401	0.2400
17	0.2217	0.2313	0.2340	0.2346	0.2368	0.2354	0.2350	0.2349
63	0.1271	0.1416	0.1460	0.1472	0.1542	0.1486	0.1478	0.1477
64	0.1261	0.1407	0.1451	0.1463	0.1534	0.1478	0.1469	0.1468
65	0.1251	0.1398	0.1442	0.1454	0.1526	0.1469	0.1460	0.1459
255	0.0544	0.0732	0.0805	0.0826	0.0820	0.0868	0.0839	0.0835
256	0.0542	0.0730	0.0803	0.0825	0.0819	0.0867	0.0837	0.0833
257		0.0728	0.0802	0.0823		0.0866	0.0836	0.0832

4. SYMMETRIC PRECONDITIONING FOR LINEAR ELLIPTIC SYSTEMS

The CGM for nonsymmetric equations in Hilbert space has been studied in Section 3.1. Using the theoretical background described in Section 2.4, superlinear convergence has been proved in Hilbert space and, based on this, mesh independence of the superlinear estimate has been derived for FEM discretizations of elliptic Dirichlet problems. The numerical realization of this method has been demonstrated in Section 3.1 for mixed elliptic problems.

Here the mesh independent superlinear convergence results are extended from a single equation to systems. First the compact normal operator approach is used for systems with homogeneous boundary conditions (cf. [36]), then we extend the results of Section 3.2 to systems using the operator pair technique (see [40]). An important advantage of the proposed preconditioning method for systems is that one can define decoupled preconditioners, hence the size of the auxiliary systems remains as small as for a single equation, moreover, parallelization of the auxiliary systems is available. The development and the numerical realization of an efficient parallel algorithm are presented at the end of this chapter, based on [39].

4.1 Systems with Dirichlet boundary conditions

4.1.1 The problem and the approach

Let us consider systems of the form

$$\left.\begin{aligned} -\operatorname{div}(K_i\,\nabla u_i) + \mathbf{b}_i\cdot\nabla u_i + \sum_{j=1}^{\ell} V_{ij}u_j &= g_i \\ u_i\big|_{\partial\Omega} &= 0 \end{aligned}\right\} \qquad (i=1,\dots,\ell) \tag{4.1}$$

satisfying the following assumptions:

Assumptions 4.1. *Suppose that*

(i) the bounded domain $\Omega\subset\mathbb{R}^d$ is C^2-diffeomorphic to a convex domain;

(ii) for all $i,j=1,\dots,\ell$ the functions $K_i\in C^1(\overline{\Omega})$, $V_{ij}\in L^\infty(\Omega)$ and $\mathbf{b}_i\in C^1(\overline{\Omega})^d$;

(iii) there exists $m > 0$ such that $K_i \geq m$ holds for all $i = 1, \ldots, \ell$;

(iv) letting $V = \{V_{ij}\}_{i,j=1}^{\ell}$, the coercivity property

$$\lambda_{\min}(V + V^T) - \max_{1 \leq i \leq \ell} \operatorname{div} \mathbf{b}_i \geq 0 \tag{4.2}$$

holds pointwise on Ω, where $\lambda_{\min}$ denotes the smallest eigenvalue;

(v) $g_i \in L^2(\Omega)$ for all $i = 1, \ldots, \ell$.

The coercivity assumption implies that problem (4.1) has a unique weak solution. Systems of the form (4.1) arise e.g. from the time discretization and Newton linearization of nonlinear reaction-convection-diffusion (transport) systems

$$\left.\begin{aligned} \frac{\partial c_i}{\partial t} - \operatorname{div}(K_i \nabla c_i) + \mathbf{b}_i \cdot \nabla c_i + R_i(x, c_1, \ldots, c_\ell) &= 0 \\ c_i\big|_{\partial\Omega} &= 0 \end{aligned}\right\} \qquad (i = 1, \ldots, \ell). \tag{4.3}$$

In many real-life problems, e.g. where c_i are concentrations of chemical species, such systems may consist of a huge number of equations (cf. [68]). Using a time discretization with sufficiently small step length τ, the systems obtained from the Newton linearization of (4.3) around some $\mathbf{c} = (c_1, \ldots, c_\ell)^T$ satisfy Assumptions 4.1. Namely, in this case

$$V(x) = \frac{\partial R(x, \mathbf{c})}{\partial \mathbf{c}} + \frac{1}{\tau}\, \mathbf{I}$$

(where $\mathbf{I}$ is the identity matrix), which ensures the coercivity (the only nontrivial assumption) for small enough τ.

For brevity, we write (4.1) as

$$\left.\begin{aligned} L\mathbf{u} \equiv -\operatorname{div}(\mathbf{K} \nabla \mathbf{u}) + \mathbf{b} \cdot \nabla \mathbf{u} + V\mathbf{u} &= \mathbf{g} \\ \mathbf{u}\big|_{\partial\Omega} &= \mathbf{0} \end{aligned}\right\} \tag{4.4}$$

where

$$\mathbf{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_\ell \end{pmatrix}, \quad \mathbf{g} = \begin{pmatrix} g_1 \\ \vdots \\ g_\ell \end{pmatrix}, \quad -\operatorname{div}(\mathbf{K} \nabla \mathbf{u}) = \begin{pmatrix} -\operatorname{div}(K_1 \nabla u_1) \\ \vdots \\ -\operatorname{div}(K_\ell \nabla u_\ell) \end{pmatrix}, \quad \mathbf{b} \cdot \nabla \mathbf{u} = \begin{pmatrix} \mathbf{b}_1 \cdot \nabla u_1 \\ \vdots \\ \mathbf{b}_\ell \cdot \nabla u_\ell \end{pmatrix}$$

and V has been defined in Assumption 4.1, condition *(iv)*. For the numerical solution of system (4.4), one usually considers its FEM discretization, which leads to a linear algebraic system

$$\mathbf{L}_h \mathbf{c} = \mathbf{g}_h. \tag{4.5}$$

Then (4.5) can be solved by the CGM using some suitable preconditioner. Here we consider preconditioners based on the following preconditioning operator. Letting $\sigma_i \in L^\infty(\Omega)$, $\sigma_i \geq 0$ be suitable functions and

$$S_i u_i := -\operatorname{div}(K_i \nabla u_i) + \sigma_i u_i \qquad (i = 1, \ldots, \ell) \tag{4.6}$$

for $u_i\big|_{\partial\Omega} = 0$, and define the ℓ-tuple of independent elliptic operators

$$S\mathbf{u} = \begin{pmatrix} S_1 u_1 \\ \vdots \\ S_\ell u_\ell \end{pmatrix}. \tag{4.7}$$

The goal of this section is twofold. First, we prove mesh independent superlinear convergence of the preconditioned CGM in the framework of compact normal operators in Hilbert space (cf. Section 2.4). This is achieved in two steps: on the theoretical level, the preconditioned form of system (4.4)

$$S^{-1}L\mathbf{u} = \mathbf{f} \equiv S^{-1}\mathbf{g} \tag{4.8}$$

will be considered and it will be proved that the CGM converges superlinearly in the Sobolev space $H_0^1(\Omega)^\ell$. Based on this, on the practically relevant discrete level we consider the preconditioned form of the algebraic system (4.5)

$$\mathbf{S}_h^{-1}\mathbf{L}_h\mathbf{c} = \mathbf{f}_h \equiv \mathbf{S}_h^{-1}\mathbf{g}_h, \tag{4.9}$$

where $\mathbf{S}_h$ denotes the discretization of S in the same FEM subspace as for $\mathbf{L}_h$, and prove that the superlinear convergence of the CGM is mesh independent, i.e. independent of the considered FEM subspace. These properties are the extension of the results of Section 3.1 to systems. On both levels the full and a truncated GCG-LS algorithms 2.28 and 2.29 are considered, and the results are proved under certain special assumptions that ensure the normality of the preconditioned operator in the corresponding Sobolev space (analogously to [9]). The second goal is the numerical testing of the proposed PCG method. Similarly to the results in Subsection 3.1.3, it turns out that the mesh independent superlinear convergence property is even valid when some of the technical conditions do not hold, i.e. beyond the normal operator framework of Section 2.4.

Besides the mesh independent convergence result, this preconditioning method has an advantage of efficient realization since the symmetric elliptic operators S_i are decoupled, hence the size of the auxiliary systems is smaller than of the original one and, moreover, parallel solution of the auxiliary systems is available. This may significantly

decrease the cost when the system (4.1) consists of many equations. This is illustrated with an example involving chemical reactions at the end of this section.

4.1.2 Iteration and convergence in Sobolev space

Let us consider the complex Hilbert space $H = L^2(\Omega)^\ell$ with inner product and corresponding norm

$$\langle \mathbf{u}, \mathbf{v} \rangle = \int_\Omega \sum_{i=1}^{\ell} u_i \overline{v}_i, \qquad \|\mathbf{u}\|^2 = \int_\Omega \sum_{i=1}^{\ell} |u_i|^2 \tag{4.10}$$

and define the operators L and S as given in (4.4) and (4.7), respectively, with the domain

$$D(L) = D(S) = D := \left(H^2(\Omega) \cap H_0^1(\Omega)\right)^\ell$$

which is dense in H. We consider problem (4.4) in H, preconditioned by S as proposed in Subsection 4.1.1. The goal is to prove Theorem 2.53 for this problem in the space $L^2(\Omega)^\ell$ by verifying that L and S satisfy Assumptions 2.48.

This will be done in two cases: first, we prove Theorem 2.53 using the truncated GCG-LS(0) algorithm 2.29 when S is the symmetric part of L. Then we consider the full GCG-LS algorithm 2.28 and prove Theorem 2.53 for problems with constant coefficients when the normality of the preconditioned operator in the corresponding Sobolev space can be ensured. This is an extension of the previous result from a single equation to systems (cf. [9]).

Remark 4.2. When the preconditioned conjugate gradient algorithms 2.51 or 2.52 are applied with L and S from (4.4) and (4.7), respectively, the auxiliary problems like $S\mathbf{z} = L\mathbf{d}$ have the following form:

$$\left.\begin{aligned} -\operatorname{div}(K_i\ \nabla z_i) + \sigma_i z_i &= L_i\mathbf{d} \\ z_i\big|_{\partial\Omega} &= 0 \end{aligned}\right\} \qquad (i = 1, \ldots, \ell),$$

where $L_i\mathbf{d} \equiv -\operatorname{div}(K_i\ \nabla d_i) + \mathbf{b}_i \cdot \nabla d_i + \sum_{j=1}^{\ell} V_{ij} d_j$ for $\mathbf{d} \in D(L)$, that is, decoupled symmetric elliptic equations have to be solved.

Convergence of the truncated algorithm

In this part we study the case when S is the symmetric part of L, i.e.

$$S = \frac{L + L^*}{2}.$$

Then the preconditioned operator $A = S^{-1}L$ has an important property in the energy space H_S (see Subsection 2.4.2). Namely, the antisymmetry of

$$Q = L - S = \frac{L - L^*}{2}$$

in H,

$$\langle Q\mathbf{u}, \mathbf{v}\rangle = -\langle \mathbf{u}, Q\mathbf{v}\rangle \tag{4.11}$$

is equivalent to the antisymmetry of $S^{-1}Q$ in H_S:

$$\left\langle S^{-1}Q\mathbf{u}, \mathbf{v}\right\rangle_S = -\left\langle \mathbf{u}, S^{-1}Q\mathbf{v}\right\rangle_S, \tag{4.12}$$

i.e. the S-adjoint operator $(S^{-1}Q)^*_S$ (cf. Remark 2.50) satisfies

$$\left(S^{-1}Q\right)^*_S = -S^{-1}Q. \tag{4.13}$$

Since $A = I + S^{-1}Q$, therefore $A^*_S = 2I - A$, hence by Remark 2.46 the truncated GCG-LS(0) version 2.52 for equation (2.31) coincides with the full algorithm 2.51.

Let us determine the symmetric part of the operator L in (4.4). We have for $\mathbf{u}, \mathbf{v} \in D$

$$\langle L\mathbf{u}, \mathbf{v}\rangle = \int_\Omega \left(\sum_{i=1}^{\ell} \left(K_i \ \nabla u_i \cdot \nabla \overline{v}_i + (\mathbf{b}_i \cdot \nabla u_i)\overline{v}_i \right) + \sum_{i,j=1}^{\ell} V_{ij} u_j \overline{v}_i \right). \tag{4.14}$$

The divergence theorem and the homogeneous Dirichlet boundary condition imply

$$\int_\Omega (\mathbf{b}_i \cdot \nabla u_i)\overline{v}_i + \int_\Omega u_i(\mathbf{b}_i \cdot \nabla \overline{v}_i) = -\int_\Omega (\operatorname{div} \mathbf{b}_i) u_i \overline{v}_i, \tag{4.15}$$

hence it is easy to see that for $\mathbf{u}, \mathbf{v} \in D$

$$\langle S\mathbf{u}, \mathbf{v}\rangle = \int_\Omega \left(\sum_{i=1}^{\ell} \left(K_i \ \nabla u_i \cdot \nabla \overline{v}_i - \frac{1}{2} \ (\operatorname{div} \mathbf{b}_i) u_i \overline{v}_i \right) + \frac{1}{2} \sum_{i,j=1}^{\ell} \left(V_{ij} + V_{ji} \right) u_j \overline{v}_i \right). \tag{4.16}$$

Hence we have coordinatewise

$$S_i\mathbf{u} = -\operatorname{div}(K_i \ \nabla u_i) - \frac{1}{2} \ (\operatorname{div} \mathbf{b}_i) u_i + \frac{1}{2} \sum_{j=1}^{\ell} (V_{ij} + V_{ji}) u_j. \tag{4.17}$$

This operator falls into the type (4.6) if and only if the antisymmetry

$$V_{ij} = -V_{ji} \qquad (i \neq j) \tag{4.18}$$

is valid and σ_i in (4.6) is chosen as

$$\sigma_i = V_{ii} - \frac{1}{2}\,(\operatorname{div} \mathbf{b}_i), \tag{4.19}$$

hence (4.18)-(4.19) are assumed to hold from now on.

As stated before, the task is to prove that the operators L and S satisfy Assumptions 2.48 in $H = L^2(\Omega)^\ell$. Together with the argument after (4.13), this will imply that the preconditioned GCG-LS(0) algorithm 2.52 converges according to Theorem 2.53.

Let us check that the conditions in Assumptions 2.48 are satisfied, when S is the symmetric part of L.

(i) S is self-adjoint by Proposition 2.12 since S_i maps onto $L^2(\Omega)$ (see [32]), hence S maps onto $L^2(\Omega)^\ell$.

(ii) Formula (4.16) yields

$$\langle S\mathbf{u}, \mathbf{u}\rangle = \int_\Omega \left(\sum_{i=1}^{\ell} \Big(K_i \;|\nabla u_i|^2 - \frac{1}{2}\,(\operatorname{div} \mathbf{b}_i)\,|u_i|^2 \Big) + \frac{1}{2} \sum_{i,j=1}^{\ell} (V_{ij} + V_{ji}) u_j \overline{u}_i \right).$$

Then conditions (iii)-(iv) in Assumptions 4.1 imply

$$\langle S\mathbf{u}, \mathbf{u}\rangle \geq m \sum_{i=1}^{\ell} \|\nabla u_i\|^2_{L^2(\Omega)}\,, \tag{4.20}$$

whence, using the Poincaré–Friedrichs inequality (2.2), letting $p = m\nu$ and using notation (4.10), we have

$$\langle S\mathbf{u}, \mathbf{u}\rangle \geq p\,\|\mathbf{u}\|^2 \qquad (\mathbf{u} \in D). \tag{4.21}$$

(iii) The antisymmetry (4.11) implies $\operatorname{Re}\langle Q\mathbf{u}, \mathbf{u}\rangle = 0$. Since $L = S + Q$, we obtain

$$\operatorname{Re}\langle L\mathbf{u}, \mathbf{u}\rangle = \langle S\mathbf{u}, \mathbf{u}\rangle\,. \tag{4.22}$$

(iv) Formula (4.16) implies that $H_S = H^1_0(\Omega)^\ell$ and the energy inner product $\langle \mathbf{u}, \mathbf{v}\rangle_S$ is the expression on the right-hand side of (4.16), equivalent to the usual one. Using (4.17)-(4.19), the antisymmetric part Q satisfies coordinatewise

$$Q_i\mathbf{u} = L_i\mathbf{u} - S_i u_i = \mathbf{b}_i \cdot \nabla u_i + \frac{1}{2}\,(\operatorname{div} \mathbf{b}_i) u_i + \sum_{\substack{j=1 \\ j\neq i}}^{\ell} V_{ij} u_j \tag{4.23}$$

for $\mathbf{u} \in (H^2(\Omega) \cap H_0^1(\Omega))^\ell$ and the same expression is valid for $\mathbf{u} \in H_0^1(\Omega)^\ell$. Then the operator $S^{-1}Q$ on $H_0^1(\Omega)^\ell$ is given by

$$\begin{aligned}\left\langle S^{-1}Q\mathbf{u}, \mathbf{v}\right\rangle_S = \langle Q\mathbf{u}, \mathbf{v}\rangle &= \int_\Omega \sum_{i=1}^{\ell} (Q_i\mathbf{u})\overline{v}_i \\ &= \int_\Omega \left(\sum_{i=1}^{\ell} \Big(\mathbf{b}_i \cdot \nabla u_i + \frac{1}{2} \, (\operatorname{div} \mathbf{b}_i) u_i \Big) \overline{v}_i + \sum_{\substack{i,j=1 \\ j \neq i}}^{\ell} V_{ij} u_j \overline{v}_i \right) \\ &\qquad \Big(\mathbf{u}, \mathbf{v} \in \big(H^2(\Omega) \cap H_0^1(\Omega) \big)^\ell \Big)\end{aligned}$$

which is compact owing to the compact embedding $H_0^1(\Omega) \hookrightarrow L^2(\Omega)$ (cf. Theorem 2.26). Further, (4.13) obviously implies that $(S^{-1}Q)_S^*$ commutes with $S^{-1}Q$, i.e. $S^{-1}Q$ is S-normal (cf. Remark 2.50).

Corollary 4.3. *Under Assumptions 4.1 and* (4.18)-(4.19)*, the preconditioned truncated GCG-LS(0) algorithm 2.52 for system* (4.1) *with the preconditioning operator* (4.6)-(4.7) *converges superlinearly in the space* $H_0^1(\Omega)^\ell$ *according to the estimate* (2.34).

In particular, in (2.34) we have the parameter $\varrho = 1$ and the norm equality $\|u\|_L = \|u\|_S$ from (4.22).

Convergence of the full algorithm

Now let us turn to the general case, when S is not the symmetric part of L, i.e. S has the form (4.6)-(4.7), but (4.18)-(4.19) are not assumed to hold. It may be important in practice to have this freedom to choose the coefficients σ_i of S. First, we have frequently $K_i = 1$ in (4.1), i.e. the term $-\operatorname{div}(K_i \, \nabla u_i)$ coincides with the Laplacian, and in such cases it may be efficient to choose σ_i constant. Namely, for auxiliary problems with constant coefficients, various fast direct solvers are available (such as fast Fourier transform or cyclic reduction, see [51, 57]) which turn S into a cheap preconditioner. Second, as shown in [45] for a single equation, large values chosen for σ may compensate for large convection terms $\mathbf{b}$, hence such a preconditioner can be useful for singularly perturbed problems as well.

As stated earlier, in order to verify Theorem 2.53 for this case, the task is to prove that the operators L and S as given in 4.4 and 4.7, respectively, satisfy Assumptions 2.48 in $H = L^2(\Omega)^\ell$. This will be proved under the restrictive condition that L has constant coefficients itself, moreover, in addition to Assumptions 4.1 the following extra properties are also assumed to hold.

Assumptions 4.4. *Suppose that*

(i) for all $i = 1, \dots, \ell$, $K_i \equiv K \in \mathbb{R}$, $\sigma_i \equiv \sigma \in \mathbb{R}$ *and* $\mathbf{b}_i \equiv \mathbf{b} \in \mathbb{R}^d$;

(ii) $V \in \mathbb{R}^{\ell \times \ell}$ *is a normal matrix.*

Then Assumptions 2.48 can be verified as follows.

(i) The same argument can be used as for the case of symmetric part preconditioning: S is self-adjoint by Proposition 2.12 since S_i maps onto $L^2(\Omega)$, hence S maps onto $L^2(\Omega)^\ell$.

(ii) Using the required form of the proposed preconditioner (4.6)-(4.7), we have

$$\langle S\mathbf{u}, \mathbf{u}\rangle = \int_\Omega \sum_{i=1}^{\ell} \left(K \ |\nabla u_i|^2 + \sigma\, |u_i|^2\right). \tag{4.24}$$

From this the assumptions $K > 0$ and $\sigma \geq 0$ imply (4.21) in the same way as it followed from (4.20).

(iii) We have for $\mathbf{u} \in D$

$$\langle L\mathbf{u}, \mathbf{u}\rangle = \int_\Omega \left(\sum_{i=1}^{\ell} \left(K \ |\nabla u_i|^2 + (\mathbf{b}\cdot\nabla u_i)\ \overline{u}_i\right) + \sum_{i,j=1}^{\ell} V_{ij} u_j \overline{u}_i \right)$$

from (4.14). Now for constant $\mathbf{b}$, (4.15) yields

$$\int_\Omega (\mathbf{b}\cdot\nabla u_i)\overline{u}_i = -\int_\Omega u_i (\mathbf{b}\cdot\nabla \overline{u}_i),$$

further, (4.2) now reduces to the assumption that $V + V^T$ is positive semidefinite. Hence

$$\begin{aligned} \operatorname{Re}\langle L\mathbf{u}, \mathbf{u}\rangle &= \int_\Omega \left(\sum_{i=1}^{\ell} \left(K \ |\nabla u_i|^2\right) + \sum_{i,j=1}^{\ell} \frac{1}{2}(V_{ij} + V_{ji}) u_j \overline{u}_i \right) \\ &\geq K \sum_{i=1}^{\ell} \|\nabla u_i\|^2_{L^2(\Omega)} + \lambda_0 \sum_{i=1}^{\ell} \|u_i\|^2_{L^2(\Omega)}, \end{aligned}$$

where

$$\lambda_0 = \lambda_{\min}\left(\frac{V_{ij} + V_{ji}}{2}\right) \geq 0.$$

Further, using (4.24) and the Poincaré–Friedrichs inequality (2.2), we obtain

$$\inf_{\substack{\mathbf{u}\in D \\ \mathbf{u}\neq\mathbf{0}}} \frac{\operatorname{Re}\langle L\mathbf{u}, \mathbf{u}\rangle}{\langle S\mathbf{u}, \mathbf{u}\rangle} \geq \inf_{\substack{(x,y)\in\mathbb{R}^2 \\ x \geq \nu y > 0}} \frac{Kx + \lambda_0 y}{Kx + \sigma y} = \min\left\{\frac{\nu K + \lambda_0}{\nu K + \sigma}, 1\right\},$$

where the latter equality comes from an elementary calculation. Therefore condition (*iii*) in Assumptions 2.48 holds with

$$\varrho = \min\left\{\frac{\nu K + \lambda_0}{\nu K + \sigma}, 1\right\}. \tag{4.25}$$

(iv) Similarly to item (iv) in the previous case, we have $H_S = H_0^1(\Omega)^\ell$ and the energy inner product $\langle \mathbf{u}, \mathbf{v}\rangle_S$ is equivalent to the usual one, further, the antisymmetric part satisfies

$$Q_i\mathbf{u} = \mathbf{b}\cdot\nabla u_i - \sigma u_i + \sum_{j=1}^{\ell} V_{ij}u_j \qquad \left(\mathbf{u}\in H_0^1(\Omega)^\ell\right), \tag{4.26}$$

whence the operator $S^{-1}Q$ on $H_0^1(\Omega)^\ell$ is compact by the same argument as for (4.23).

On the other hand, the normality of $S^{-1}Q$ in H_S is not as trivial as in the previous subsection (since it is not antisymmetric), but this is the main property to be verified now in two steps.

Lemma 4.5. *Let us define the operators $R, W : L^2(\Omega)^\ell \to L^2(\Omega)^\ell$ by*

$$\begin{aligned} R\mathbf{u} &:= \left(\mathbf{b}\cdot\nabla u_i\right)_{i=1}^{\ell} \quad \left(\mathbf{u}\in D(R) = H_0^1(\Omega)^\ell\right) \\ W\mathbf{u} &:= V\mathbf{u} - \sigma\mathbf{u} \quad \left(\mathbf{u}\in L^2(\Omega)^\ell\right), \end{aligned} \tag{4.27}$$

respectively. Then the following operators commute:

(a) $S^{-1}W$ and $S^{-1}W^*$;

(b) $S^{-1}R$ and $S^{-1}W$;

(c) $S^{-1}R$ and $S^{-1}W^*$.

Proof. First we observe

$$SW\mathbf{u} = WS\mathbf{u} \qquad (\mathbf{u}\in D(S)) \tag{4.28}$$

since, using $S\mathbf{u} = -K\ \operatorname{diag}(\Delta u_i)$, (4.28) is coordinatewise equivalent to

$$\Delta\left(\sum_{j=1}^{\ell} W_{ij}u_j\right) = \left(\sum_{j=1}^{\ell} W_{ij}\Delta u_j\right).$$

Replacing $\mathbf{u}$ by $S^{-1}\mathbf{u}$ in (4.28) (which makes sense since S maps onto $L^2(\Omega)^\ell$) and

applying S^{-1} to both sides, we obtain

$$WS^{-1}\mathbf{u} = S^{-1}W\mathbf{u} \qquad \left(\mathbf{u} \in L^2(\Omega)^\ell\right) \tag{4.29}$$

(a) Using (4.29) and its analogue for W^*, further that W is normal (inheriting this from V), we obtain

$$WS^{-1}W^* = S^{-1}WW^* = S^{-1}W^*W = W^*S^{-1}W.$$

Applying S^{-1} to both sides we obtain the required statement.

(b) Introducing the operators $S_0 := -K\Delta$ and $R_0 := \mathbf{b}\cdot\nabla$, we have $S = \operatorname{diag}(S_0)$ and $R = \operatorname{diag}(R_0)$. Using that these operators have constant coefficients, one can prove $R_0S_0^{-1} = S_0^{-1}R_0$ (see [9, Prop. 1]), therefore we obtain $RS^{-1} = S^{-1}R$. We have $RW = WR$ similarly to (4.28), and using also (4.29) we obtain

$$RS^{-1}W = S^{-1}RW = S^{-1}WR = WS^{-1}R.$$

Applying S^{-1} to both sides again, we obtain the required statement.

(c) This follows from (b) by replacing W by W^*. □

Proposition 4.6. *The operator $S^{-1}Q$ is normal in H_S.*

Proof. Relations (4.26) and (4.27) imply $Q = R + W$, hence

$$S^{-1}Q = S^{-1}R + S^{-1}W. \tag{4.30}$$

Here the S-adjoints of the operators on the right-hand side are as follows. First, now for constant $\mathbf{b}$ the equality (4.15) implies for all $\mathbf{u}, \mathbf{v} \in H_0^1(\Omega)^\ell$

$$\langle R\mathbf{u}, \mathbf{v}\rangle = -\langle \mathbf{u}, R\mathbf{v}\rangle,$$

that is,

$$\left\langle S^{-1}R\mathbf{u}, \mathbf{v}\right\rangle_S = -\left\langle \mathbf{u}, S^{-1}R\mathbf{v}\right\rangle_S$$

which means that $(S^{-1}R)_S^* = -S^{-1}R$. Further,

$$\left\langle S^{-1}W\mathbf{u}, \mathbf{v}\right\rangle_S = \langle W\mathbf{u}, \mathbf{v}\rangle = \langle \mathbf{u}, W^*\mathbf{v}\rangle = \left\langle \mathbf{u}, S^{-1}W^*\mathbf{v}\right\rangle_S,$$

i.e. $(S^{-1}W)^*_S = S^{-1}W^*$. Altogether, we have

$$\left(S^{-1}Q\right)^*_S = -S^{-1}R + S^{-1}W^*,$$

which by (4.30) and Lemma 4.5 commutes with $S^{-1}Q$. □

Corollary 4.7. *Under Assumptions 4.1 and 4.4, the preconditioned full GCG-LS algorithm 2.51 for system* (4.1) *with the preconditioning operator* (4.6)-(4.7) *converges superlinearly in the space* $H^1_0(\Omega)^\ell$ *according to the estimate* (2.34).

In particular, we have the expression (4.25) for the parameter ϱ in (2.34).

4.1.3 Mesh independent superlinear convergence for the discretized problem

In this section we derive the main result from practical point of view. Let us consider the FEM discretization of system (4.4) in some FEM subspace

$$V_h = \operatorname{span}\{\varphi_1, \varphi_2, \ldots, \varphi_n\} \subset H^1_0(\Omega)^\ell,$$

which leads to an $n \times n$ linear algebraic system

$$\mathbf{L}_h\mathbf{c} = \mathbf{g}_h. \tag{4.31}$$

Let $\mathbf{S}_h$ denote the discretization of S in the same FEM subspace V_h as for $\mathbf{L}_h$. We consider the preconditioned form of the algebraic system (4.31)

$$\mathbf{S}_h^{-1}\mathbf{L}_h\mathbf{c} = \mathbf{f}_h \equiv \mathbf{S}_h^{-1}\mathbf{g}. \tag{4.32}$$

Here we show that the superlinear convergence of the CGM is mesh independent, i.e. independent of the subspace V_h. Namely, by Section 4.1.2, under the given conditions, the operators L and S as given in (4.4) and (4.7), respectively, satisfy Assumptions 2.48 for the operator equation (2.29). Let $V_h \subset H_S$ be a finite dimensional subspace, $\mathbf{S}_h$ and $\mathbf{Q}_h$ the corresponding Gram matrices of S and Q, respectively. If the matrix $\mathbf{S}_h^{-1}\mathbf{Q}_h$ is $\mathbf{S}_h$-normal, then the conditions of Theorem 2.54 are satisfied and the mesh independent superlinear convergence estimate (2.37) holds.

Consequently, mesh independence result for the elliptic system (4.1) is obtained under the conditions considered in Subsection 4.1.2 to verify Assumptions 2.48. To formulate this, we note that with symmetric part preconditioning, the $\mathbf{S}_h$-normality of the matrix $\mathbf{S}_h^{-1}\mathbf{Q}_h$ need not be assumed since it holds for an arbitrary FEM subspace (cf. Remark 2.55).

Corollary 4.8. *Let Assumptions 4.1 hold. Consider the FEM discretization of system* (4.1)*, using the stiffness matrix of* (4.7) *as preconditioner, under one of the following conditions:*

(a) properties (4.18)-(4.19) *hold, $V_h \subset H_0^1(\Omega)^\ell$ is an arbitrary FEM subspace and the truncated GCG-LS(0) algorithm 2.52 is used (here the $\mathbf{S}_h$-normality of $\mathbf{S}_h^{-1}\mathbf{Q}_h$ automatically holds);*

(b) Assumptions (4.4) *hold, $V_h \subset H_0^1(\Omega)^\ell$ is a FEM subspace for which the matrix $\mathbf{S}_h^{-1}\mathbf{Q}_h$ is $\mathbf{S}_h$-normal, and the full GCG-LS 2.51 is used.*

Then the mesh independent superlinear convergence estimate (2.37) *is valid.*

If symmetric part preconditioning is used, that is, the conditions in item (a) hold, then estimate (2.37) holds with $\varrho = 1$ and the error is measured in $\mathbf{S}_h$-norm.

Remark 4.9. Following Remark 4.2, the CGM for system (4.32) involves the FEM solution of decoupled Helmholtz problems of the following type in the subspace V_h:

$$\left.\begin{aligned} -\operatorname{div}(K_i\,\nabla z_i) + \sigma_i z_i &= L_i\mathbf{d} \\ z_i\big|_{\partial\Omega} &= 0 \end{aligned}\right\} \qquad (i = 1, \ldots, \ell),$$

This provides the following advantages for the studied PCG algorithm:

- the size of the auxiliary systems is considerably smaller than that of the original system when ℓ is large;
- parallel solution of the auxiliary systems is available;
- for Helmholtz preconditioners various efficient solvers are available (like fast Fourier transform, cyclic reduction or multigrid, see e.g. [25, 51, 57]).

4.1.4 Numerical experiments

In this subsection some numerical results are presented. Besides illustrating the preceding theorems, the main outcome of this test is that the mesh independent superlinear convergence property is even valid when some of the previous theoretical conditions do not hold. This means that the normal operator framework of Section 2.4 seems to be only technical, although currently the obstacles are deemed to be insuperable. Consequently, the proposed preconditioned CGM is an efficient solution method for general elliptic problems.

In what follows, let $\Omega \subset \mathbb{R}^2$ be the unit square and $K_i = 1$ $(i = 1, \ldots, \ell)$ in (4.1), i.e. for simplicity only the case of Laplacian is considered for the principal part of the

elliptic operators. Since in this subsection only Dirichlet boundary conditions $u_i\big|_{\partial\Omega} = 0$ are investigated, the indication of the boundary conditions will be omitted. Both of the studied algorithms will be used: the truncated one where possible and the full algorithm throughout.

In the first part of this subsection, systems consisting of 2 and 3 equations are investigated. In both cases we consider a system that does and one that does not satisfy the theoretical conditions. Finally we consider a larger model involving chemical reactions between 10 pollutants. The numbers in the tables are the values of

$$Q_k := \left(\frac{\|e_k\|_{\mathbf{L}_h}}{\|e_0\|_{\mathbf{L}_h}} \right)^{1/k}$$

for the iteration counter parameter $k = 1, 2, \ldots$ In all the experiments numerical superlinear convergence has been observed (i.e. that Q_k decreases) up to some point when this decrease has stopped. Here we usually had

$$\frac{\|e_k\|_{\mathbf{L}_h}}{\|e_0\|_{\mathbf{L}_h}} \approx 10^{-14},$$

which has justified stopping the iteration.

Experiment 1 Let

$$\mathbf{b} \equiv \mathbf{b}_i = (1, 0), \qquad V = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

The system according to these parameters is the following:

$$\left.\begin{aligned} -\Delta u_1 + \partial_x u_1 + u_2 &= g_1, \\ -\Delta u_2 + \partial_x u_2 - u_1 &= g_2. \end{aligned}\right\} \tag{4.33}$$

Here the truncated algorithm is also applicable since $V_{ij} = -V_{ji}$ $(i, j = 1, 2)$. If we choose $\sigma_i = V_{ii} = 0$ $(i = 1, \ldots, \ell)$, i.e. the preconditioner is $S_i u_i = -\Delta u_i$ (see the conditions (4.18)-(4.19)), then the corresponding full GCG-LS algorithm coincides with the truncated version.

The values of Q_k in the columns of Table 4.1 show the superlinear convergence, moreover, the rows show the boundedness of Q_k as the mesh parameter increases. Larger values of σ_i can also improve the convergence of the full algorithm, although the computational cost is increased by assembling the mass matrix. Similar results can be found in Table 4.2 for the second experiment.

Tab. 4.1: Values of Q_k for system (4.33).

	1/h					
	truncated algorithm			full algorithm, $\sigma_i = 8$		
Itr.	32	64	128	32	64	128
1	0.0774	0.0776	0.0776	0.0636	0.0638	0.0638
2	0.0777	0.0780	0.0780	0.0624	0.0626	0.0626
3	0.0802	0.0805	0.0805	0.0642	0.0644	0.0644
4	0.0777	0.0780	0.0781	0.0643	0.0645	0.0646
5	0.0720	0.0723	0.0724	0.0616	0.0618	0.0619
6	0.0663	0.0666	0.0667	0.0579	0.0581	0.0582
7	0.0617	0.0620	0.0621	0.0542	0.0545	0.0546
8	0.0587	0.0590	0.0590	0.0511	0.0514	0.0515
9	0.0574	0.0576	0.0577	0.0489	0.0491	0.0491
10	0.0564	0.0567	0.0568	0.0482	0.0483	0.0483

Experiment 2 Let

$$\mathbf{b}_1 = (1,0), \quad \mathbf{b}_2 = (0,1), \qquad V = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},$$

in other words we have

$$\left.\begin{aligned} -\Delta u_1 + \partial_x u_1 + u_2 &= g_1 \\ -\Delta u_2 + \partial_y u_2 - u_1 &= g_2. \end{aligned}\right\} \tag{4.34}$$

Tab. 4.2: Values of Q_k for system (4.34).

	1/h					
	truncated algorithm			full algorithm, $\sigma_i = 8$		
Itr.	32	64	128	32	64	128
1	0.0851	0.0853	0.0854	0.0671	0.0672	0.0672
2	0.0838	0.0841	0.0841	0.0678	0.0680	0.0680
3	0.0762	0.0766	0.0766	0.0638	0.0641	0.0641
4	0.0705	0.0709	0.0709	0.0598	0.0601	0.0601
5	0.0675	0.0678	0.0678	0.0566	0.0569	0.0570
6	0.0665	0.0668	0.0668	0.0548	0.0551	0.0552
7	0.0656	0.0659	0.0660	0.0545	0.0547	0.0547
8	0.0633	0.0637	0.0638	0.0545	0.0547	0.0548
9	0.0600	0.0605	0.0606	0.0532	0.0535	0.0536
10	0.0569	0.0574	0.0576	0.0510	0.0514	0.0515

Experiment 3 Let

$$\mathbf{b} = \mathbf{b}_i = (1,0), \qquad V = \begin{pmatrix} 2 & 1 & 0 \\ -1 & 2 & -1 \\ 0 & 1 & 2 \end{pmatrix},$$

Thus we have following system:

$$\left.\begin{aligned} -\Delta u_1 + \partial_x u_1 + 2u_1 + u_2 &= g_1 \\ -\Delta u_2 + \partial_x u_2 - u_1 + 2u_2 - u_3 &= g_2 \\ -\Delta u_3 + \partial_x u_3 + u_2 + 2u_3 &= g_2 \end{aligned}\right\} \tag{4.35}$$

Here the truncated algorithm is again applicable. Since $V_{ii} = 2$, the truncated and the full algorithms provide the same result when $\sigma_i = 2$ is chosen in the preconditioner S_i. Table 4.3 shows the results for both algorithms.

Tab. 4.3: Values of Q_k for system (4.35).

	1/h							
	truncated alg.		full alg., $\sigma_i = 0$		full alg., $\sigma_i = 2$		full alg., $\sigma_i = 8$	
Itr.	32	128	32	128	32	128	32	128
1	0.0860	0.0863	0.0910	0.0913	0.0860	0.0863	0.0741	0.0743
2	0.0834	0.0837	0.0878	0.0882	0.0834	0.0837	0.0729	0.0731
3	0.0816	0.0819	0.0861	0.0865	0.0816	0.0819	0.0713	0.0716
4	0.0804	0.0807	0.0853	0.0856	0.0804	0.0807	0.0697	0.0699
5	0.0779	0.0782	0.0823	0.0827	0.0779	0.0782	0.0676	0.0678
6	0.0742	0.0745	0.0778	0.0782	0.0742	0.0745	0.0652	0.0655
7	0.0697	0.0701	0.0725	0.0730	0.0697	0.0701	0.0624	0.0627
8	0.0657	0.0661	0.0681	0.0686	0.0657	0.0661	0.0595	0.0598
9	0.0628	0.0633	0.0652	0.0657	0.0628	0.0633	0.0570	0.0574
10	0.0612	0.0617	0.0635	0.0640	0.0612	0.0617	0.0555	0.0559

Experiment 4 Let

$$\mathbf{b}_1 = (1,0), \quad \mathbf{b}_2 = (0,1), \quad \mathbf{b}_3 = (2,-1), \qquad V = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 2 & 1 \\ 0 & 0 & -3 \end{pmatrix}, \tag{4.36}$$

In (4.36) the matrix V does not satisfy the antisymmetry relation (4.18), it is not even normal, the coercivity property (4.2) does not hold (because of the presence of negative eigenvalues of $V + V^T$) and every $\mathbf{b}_i$ is different.

Tab. 4.4: Values of Q_k for system (4.36).

	1/h					
	full alg., $\sigma_i = 0$		full alg., $\sigma_i = 2$		full alg., $\sigma_i = 8$	
Itr.	32	128	32	128	32	128
1	0.1685	0.1689	0.1595	0.1598	0.1376	0.1379
2	0.1626	0.1630	0.1549	0.1553	0.1359	0.1362
3	0.1485	0.1489	0.1429	0.1434	0.1285	0.1288
4	0.1360	0.1365	0.1318	0.1323	0.1208	0.1212
5	0.1254	0.1261	0.1222	0.1229	0.1136	0.1141
6	0.1175	0.1182	0.1147	0.1154	0.1073	0.1079
7	0.1107	0.1114	0.1081	0.1088	0.1015	0.1022
8	0.1042	0.1050	0.1019	0.1026	0.0962	0.0969
9	0.0985	0.0993	0.0966	0.0973	0.0917	0.0924
10	0.0946	0.0954	0.0929	0.0937	0.0885	0.0892
11	0.0915	0.0924	0.0900	0.0908	0.0857	0.0864
12	0.0887	0.0895	0.0871	0.0879	0.0828	0.0836

Here the symmetric part of the equation does not provide decoupled preconditioners, thus only Algorithm 2.51 was used. Table 4.4 shows that the algorithm still has the superlinear property in spite of the fact that none of the required conditions are valid. Although the numbers Q_k are larger, the level of decreasing is approximately the same.

Experiment 5 Now let us consider a more realistic problem. The following system of equations comes from a simplified meteorological model after time discretization and linearization, based on [68]. We have

$$V = \begin{pmatrix} 0 & k_5 & 0 & 0 & -k_6 & 0 & -k_4 & -k_3 & 0 & 0 \\ 0 & -k_5 & 0 & 0 & k_6 & 0 & k_4 & k_3 & -k_9 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -k_1 & 0 \\ 0 & 0 & 0 & -k_2 & 0 & 0 & 0 & k_3 & 2k_1 & 0 \\ 0 & k_5 & 0 & 0 & -k_6 & 0 & 0 & -k_8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & k_9 & 0 \\ 0 & 0 & 0 & 2k_2 & 0 & 0 & -k_4 & k_3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -k_3 & 4k_1 & 0 \\ 0 & -k_9 & -k_6 & 0 & 0 & 0 & k_4 + 2k_8 & 0 & 0 & 0 \\ 0 & -k_8 & 0 & 0 & k_7 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \tag{4.37}$$

where the coefficients k_i can be determined from chemical reactions, see Table 4.5. Further, we have $\mathbf{b}_i = (1/10, 0)$ and the right-hand sides of the equations come from the results from the previous time-step.

The time step $\tau = 0.2829 \cdot 10^{-3}$ was chosen sufficiently small to ensure the coercivity

Tab. 4.5: The coefficients of the chemical reactions.

k_1	$6.00 \cdot 10^{-12}$	$1.60 \cdot 10^{-14}$	k_6
k_2	$7.80 \cdot 10^{-05}$	$1.90 \cdot 10^{-04}$	k_7
k_3	$8.00 \cdot 10^{-12}$	$2.30 \cdot 10^{-10}$	k_8
k_4	$8.00 \cdot 10^{-12}$	$1.00 \cdot 10^{-11}$	k_9
k_5	$1.00 \cdot 10^{-02}$	$2.90 \cdot 10^{-13}$	k_{10}

property. Further, for suitable balancing different coefficients σ_i were chosen, namely:

$$\sigma = \tau \cdot \left(1,\ 100,\ 1,\ 10,\ 1,\ 1,\ 1,\ 1,\ \frac{1}{10},\ \frac{1}{100}\right).$$

In this experiment the time of computing has also been measured: since this system consists of ten equations, the iteration with solving only block-diagonal symmetric auxiliary problems is expectedly faster than the direct solution with the nonsymmetric full matrix.

Tab. 4.6: Values of Q_k for the chemical system.

	1/h			
Itr.	8	16	32	64
1	0.0073	0.0076	0.0076	0.0077
2	0.0067	0.0071	0.0072	0.0072
3	0.0060	0.0065	0.0066	0.0066
4	0.0054	0.0060	0.0061	0.0061
5	0.0048	0.0054	0.0056	0.0056
6	0.0043	0.0050	0.0052	0.0053

In the first phase of the algorithm the matrices $\mathbf{S}_h$ and $\mathbf{Q}_h$ are constructed. The direct solution requires solving the nonsymmetric linear algebraic system

$$\mathbf{L}_h\mathbf{c} \equiv (\mathbf{S}_h + \mathbf{Q}_h)\mathbf{c} = \mathbf{g}_h.$$

The iterative algorithm solves equations like $\mathbf{S}_h z_h = d_h$ as many times as many iteration step is chosen. Here the auxiliary equations were solved by using the Cholesky decomposition of $\mathbf{S}_h$.

The run-times for this system can be found in Table 4.7. The last two colums show the difference between the direct solution and the preconditioned conjugate gradient method. The numbers in the last column are the total time of the decomposition and the iteration. It also shows that the CGM with suitable decoupled preconditioners provides better results even for mid-sized problems.

Tab. 4.7: Computational time

1/h	$\mathbf{S}_h$, $\mathbf{L}_h$	Cholesky	iteration	direct solution	PCG
8	0.0470	0.0470	0.5780	0.0150	0.6250
16	0.1090	0.0620	1.2350	0.3130	1.2970
32	0.4220	0.1880	3.9680	9.5780	4.1560
64	1.9070	2.3600	17.8120	177.7030	20.1720

4.2 Systems with nonhomogeneous mixed boundary conditions

The results of the previous section can be generalized further for systems with homogeneous mixed boundary conditions, using the operators in weak form and the weakly defined symmetric part of Subsection 2.4.2. Moreover, it has turned out from Section 3.2 that nonhomogeneous mixed boundary conditions cause no difficulties, they can be handled by using operator pairs. Here we sum up briefly the results of Section 3.2 for systems where the preconditioners are chosen to be decoupled as in Section 4.1.

Let us consider elliptic systems of the form

$$\left.\begin{aligned} -\operatorname{div}(A_i\,\nabla u_i) + \mathbf{b}_i\cdot\nabla u_i + \sum_{j=1}^{\ell} V_{ij}u_j &= g_i \\ \frac{\partial u_i}{\partial \nu_{A_i}} + \alpha_i u_i\big|_{\Gamma_N} &= \gamma_i \\ u_i\big|_{\Gamma_D} &= 0 \end{aligned}\right\} \qquad (i=1,\dots,\ell) \tag{4.38}$$

satisfying the combination of Assumptions 3.5 and 4.1:

Assumptions 4.10. *Suppose that*

(i) $\Omega\subset\mathbb{R}^d$ *is a bounded piecewise* C^1 *domain;* Γ_D,Γ_N *are disjoint open measurable subparts of* $\partial\Omega$ *such that* $\partial\Omega=\overline{\Gamma}_D\cup\overline{\Gamma}_N$*;*

(ii) for all $i,j=1,\dots,\ell$ *the matrix-valued functions* $A_i\in L^\infty(\overline{\Omega},\mathbb{R}^{d\times d})$ *and for all* $x\in\overline{\Omega}$ *the matrices* $A_i(x)$ *are symmetric; further,* $\mathbf{b}_i\in W^{1,\infty}(\Omega)^d$*,* $V_{ij}\in L^\infty(\Omega)$ *and* $\alpha_i\in L^\infty(\Gamma_N)$*;*

(iii) There exists $p>0$ *such that* $A_i(x)\xi\cdot\xi\ge p\,|\xi|^2$ *for all* $x\in\overline{\Omega}$*,* $\xi\in\mathbb{R}^d$ *and for any* $i=1,\dots,\ell$*;*

(iv) letting $V=(V_{ij})_{i,j=1}^{\ell}$*, the coercivity property*

$$\hat{c}:=\lambda_{\min}(V+V^T)-\max_{1\le i\le\ell}\operatorname{div}\mathbf{b}_i\ge 0 \tag{4.39}$$

holds pointwise on Ω*, where* $\lambda_{\min}$ *denotes the smallest eigenvalue, and*

$$\hat{\alpha}_i := \alpha_i + \frac{1}{2}\left(\mathbf{b}_i \cdot \nu\right) \geq 0 \tag{4.40}$$

holds on Γ_N *for any* $i = 1, \ldots, \ell$*;*

(v) $g_i \in L^2(\Omega), \gamma_i \in L^2(\Gamma_N)$ *for all* $i = 1, \ldots, \ell$*;*

(vi) either $\Gamma_D \neq \emptyset$*, or* $\hat{c}$ *or* $\min\limits_{1 \leq i \leq \ell} \hat{\alpha}_i$ *has a positive lower bound.*

These assumptions imply that problem (4.38) has a unique weak solution. For brevity, we write (4.38) as

$$\left.\begin{aligned} -\operatorname{div}(\mathbf{A}\ \nabla \mathbf{u}) + \mathbf{b} \cdot \nabla \mathbf{u} + V\mathbf{u} &= \mathbf{g} \\ \mathbf{u}\big|_{\Gamma_D} = \mathbf{0}, \qquad \frac{\partial \mathbf{u}}{\partial \nu_{\mathbf{A}}} + \boldsymbol{\alpha}\mathbf{u}\big|_{\Gamma_N} &= \boldsymbol{\gamma} \end{aligned}\right\} \tag{4.41}$$

The equivalent operator approach can be extended to systems, where the corresponding operator L is defined as an ℓ-tuple of operator pairs:

$$L = (L_1, \ldots, L_\ell) = \left(\begin{pmatrix} M_1 \\ P_1 \end{pmatrix}, \ldots, \begin{pmatrix} M_\ell \\ P_\ell \end{pmatrix}\right), \tag{4.42}$$

where

$$L_i \equiv \begin{pmatrix} M_i \\ P_i \end{pmatrix}, \quad L_i \begin{pmatrix} \mathbf{u} \\ \eta_i \end{pmatrix} = \begin{pmatrix} M_i \mathbf{u} \\ P_i \eta_i \end{pmatrix} = \begin{pmatrix} -\operatorname{div}(A_i\ \nabla u_i) + \mathbf{b}_i \cdot \nabla u_i + (V\mathbf{u})_i \\ \dfrac{\partial \eta_i}{\partial \nu_{A_i}} + \alpha_i \eta_i\big|_{\Gamma_N} \end{pmatrix}. \tag{4.43}$$

Using the notations of Subsection 3.2.1 and the preconditioning approach of Subsection 4.1.1, one can define the preconditioning operator

$$S = (S_1, \ldots, S_\ell) = \left(\begin{pmatrix} N_1 \\ R_1 \end{pmatrix}, \ldots, \begin{pmatrix} N_\ell \\ R_\ell \end{pmatrix}\right) \tag{4.44}$$

as the ℓ-tuple of independent operators

$$S_i \equiv \begin{pmatrix} N_i \\ R_i \end{pmatrix}, \quad S_i \begin{pmatrix} u_i \\ \eta_i \end{pmatrix} = \begin{pmatrix} N_i u_i \\ R_i \eta_i \end{pmatrix} = \begin{pmatrix} -\operatorname{div}(G_i\ \nabla u_i) + \sigma_i u_i \\ \dfrac{\partial \eta_i}{\partial \nu_{G_i}} + \beta_i \eta_i\big|_{\Gamma_N} \end{pmatrix} \tag{4.45}$$

satisfying similar assumptions as of L:

Assumptions 4.11. *Suppose that (for all* $i = 1, \ldots, \ell$*)*

(i) substituting G_i *for* A_i*,* Ω*,* Γ_D*,* Γ_N *and* G_i *satisfy Assumptions 4.10;*

(ii) $\sigma_i \in L^\infty(\Omega)$, $\sigma_i \geq 0$, $\beta_i \in L^\infty(\Gamma_N)$, $\beta_i \geq 0$*; further, if* $\Gamma_D \neq \emptyset$*, then* $\min\limits_{1\leq i\leq \ell} \sigma_i$ *or* $\min\limits_{1\leq i\leq \ell} \beta_i$ *has a positive lower bound.*

As in (3.27) for a single equation, here we look for the weak solution of the operator equation

$$L \begin{pmatrix} \mathbf{u} \\ \mathbf{u}\big|_{\Gamma_N} \end{pmatrix} = \begin{pmatrix} \mathbf{g} \\ \boldsymbol{\gamma} \end{pmatrix}. \tag{4.46}$$

If $V_h \subset H_D^1(\Omega)$ is a finite dimensional FEM subspace, then the discretization of (4.38) in V_h^ℓ leads to a linear algebraic system

$$\mathbf{L}_h \mathbf{c} = \mathbf{d}_h. \tag{4.47}$$

Let us take the symmetric operator given in (4.44)-(4.45) and introduce the corresponding stiffness matrix $\mathbf{S}_h$ in $H_D^1(\Omega)^\ell$. Then the preconditioned form of (4.47) becomes

$$\mathbf{S}_h^{-1}\mathbf{L}_h \mathbf{c} = \left(\mathbf{I}_h + \mathbf{S}_h^{-1}\mathbf{Q}_h\right) \mathbf{c} = \mathbf{S}_h^{-1}\mathbf{d}_h, \tag{4.48}$$

where $\mathbf{Q}_h = \mathbf{L}_h - \mathbf{S}_h$. Extending the results of Subsection 3.2.1 to systems, it is easy to verify that for $G_i = A_i$ $(i = 1, \ldots, \ell)$ the operators L and S are compact-equivalent with $\mu = 1$, i.e.

$$L_S = I + Q_S$$

holds in H_S with some compact operator Q_S. The energy inner product has the form

$$\begin{aligned} \left\langle \begin{pmatrix} \mathbf{u} \\ \mathbf{u}\big|_{\Gamma_N} \end{pmatrix}, \begin{pmatrix} \mathbf{v} \\ \mathbf{v}\big|_{\Gamma_N} \end{pmatrix} \right\rangle_S &= \left\langle S \begin{pmatrix} \mathbf{u} \\ \mathbf{u}\big|_{\Gamma_N} \end{pmatrix}, \begin{pmatrix} \mathbf{v} \\ \mathbf{v}\big|_{\Gamma_N} \end{pmatrix} \right\rangle_{H^\ell} \\ &= \sum_{i=1}^{\ell} \left(\langle N_i u_i, v_i \rangle_{L^2(\Omega)} + \left\langle R_i u_i\big|_{\Gamma_N}, v_i\big|_{\Gamma_N} \right\rangle_{L^2(\Gamma_N)} \right) \\ &= \int_\Omega \left[\sum_{i=1}^{\ell} \left(G_i \, \nabla u_i \cdot \nabla v_i + \sigma_i u_i v_i \right) \right] + \int_{\Gamma_N} \sum_{i=1}^{\ell} \beta_i u_i v_i. \end{aligned} \tag{4.49}$$

Similarly to Section 3.2, Green's formula implies that

$$\begin{aligned} \left\langle L_S \begin{pmatrix} \mathbf{u} \\ \mathbf{u}\big|_{\Gamma_N} \end{pmatrix}, \begin{pmatrix} \mathbf{v} \\ \mathbf{v}\big|_{\Gamma_N} \end{pmatrix} \right\rangle_S &= \sum_{i=1}^{\ell} \left(\langle M_i \mathbf{u}, v_i \rangle_{L^2(\Omega)} + \left\langle P_i u_i\big|_{\Gamma_N}, v_i\big|_{\Gamma_N} \right\rangle_{L^2(\Gamma_N)} \right) \\ &= \int_\Omega \left[\sum_{i=1}^{\ell} \left(A_i \, \nabla u_i \cdot \nabla v_i + (\mathbf{b}_i \cdot \nabla u_i) \, v_i + \sum_{j=1}^{\ell} V_{ij} u_j v_i \right) \right] + \int_{\Gamma_N} \sum_{i=1}^{\ell} \alpha_i u_i v_i. \end{aligned}$$

Analogously to the construction of (3.40), the symmetric part of L_S has the form

$$\int_\Omega \left[\sum_{i=1}^{\ell} \left(A_i \, \nabla u_i \cdot \nabla v_i - \frac{1}{2} \left(\operatorname{div} \mathbf{b}_i \right) u_i v_i \right) + \frac{1}{2} \sum_{i,j=1}^{\ell} \left(V_{ij} + V_{ji} \right) u_i v_j \right] + \int_{\Gamma_N} \sum_{i=1}^{\ell} \left(\alpha_i + \frac{1}{2} \left(\mathbf{b}_i \cdot \nu \right) \right) u_i v_i,$$

which falls into the type of (4.49) if and only if $G_i = A_i$ and

$$V_{ij} = -V_{ji} \quad (i \neq j), \qquad \sigma_i = V_{ii} - \frac{1}{2} (\operatorname{div} \mathbf{b}_i), \qquad \beta_i = \hat{\alpha}_i \equiv \alpha_i + \frac{1}{2} \left(\mathbf{b}_i \cdot \nu \right). \quad (4.50)$$

Now let us consider the preconditioned equation (4.48), when $\mathbf{L}_h$ and $\mathbf{S}_h$ now come from the elliptic operators L and S, $\mathbf{Q}_h = \mathbf{L}_h - \mathbf{S}_h$. When symmetric part preconditioning is used, that is, the preconditioner S is defined as in (4.45) with the conditions (4.50), then $Q_S \in B\left(H_S\right)$, which is now the sum of bilinear forms that can be constructed analogously to (3.41), is a compact normal operator and the matrix $\mathbf{S}_h^{-1}\mathbf{Q}_h$ is $\mathbf{S}_h$-normal with respect to $\langle \cdot, \cdot \rangle_{\mathbf{S}_h}$. In this case the superlinear convergence estimate (3.13) holds, and the GCG-LS method reduces to the truncated GCG-LS(0) algorithm 2.29.

When S is not the symmetric part of L, then $Q_S \in B\left(H_S\right)$ can be defined as the sum of similar operators corresponding to (3.42). Now the conditions of Theorem 2.44 are satisfied, thus the CGN algorithm 2.33 provides a similar mesh independent superlinear convergence result.

Corollary 4.12. *With Assumptions 4.10-4.11 and $A_i = G_i$ $(i = 1, \ldots, \ell)$, the CGN algorithm 2.33 for system* (4.48) *yields*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}} \right)^{1/k} \leq \frac{2}{m^2} \left(\frac{1}{k} \sum_{j=1}^{k} \left(\, |\lambda_j(Q_S^* + Q_S)| + \lambda_j(Q_S^* Q_S) \right) \right) \xrightarrow{k \to \infty} 0,$$

where $m > 0$ comes from the S-coercivity of L in Proposition 3.7.

The main advantage of the preconditioner (4.45) is that the corresponding stiffness matrix is block diagonal. This means that ℓ independent auxiliary linear systems have to be solved in the CGN algorithm (twice in each iteration step), as explained in Remark 3.11 for the GCG-LS algorithms.

4.3 A parallel algorithm for decoupled preconditioners

Let us return to system (4.1) and apply the preconditioned (full or truncated) GCG-LS algorithm. As it has turned out from Section 4.1, using the proposed preconditioner

(4.6)-(4.7) one has to solve auxiliary decoupled elliptic problems

$$\left.\begin{aligned} -\operatorname{div}(K_i\ \nabla z_i) + \sigma_i z_i &= L_i \mathbf{d} \\ z_i\big|_{\partial\Omega} &= 0 \end{aligned}\right\} \qquad (i = 1, \ldots, \ell), \tag{4.51}$$

in the FEM subspace V_h. The main advantage is that the stiffness matrix of the proposed preconditioner is block diagonal, hence the size of the auxiliary systems is much smaller than the size of the original system. Moreover, fast solvers are available for Helmholtz problems, i.e. for constant coefficients in (4.51). In this section a parallel algorithm is developed and tested on a multiprocessor architecture.

4.3.1 Parallelization of the GCG-LS algorithm

The basic advantage of the proposed preconditioner is its inherent parallelism. The kth iteration of the full version of the preconditioned GCG-LS algorithm 2.51 consists of two matrix-vector multiplications with matrix $\mathbf{L}_h$, one preconditioning step (solving a system of equations with the preconditioner), solving a system of s_k equations, $3s_k+2$ inner products, and $s_k + 2$ linked triads (a vector updated by a vector multiplied by a scalar).

Let us consider a parallel system with p processors. We divide the vectors u_k, d_k, r_k, z_k (defined in Algorithm 2.51) in such a way that the first $\left\lceil \frac{\ell}{p} \right\rceil$ blocks are stored in the first processor, blocks for $i = \left\lceil \frac{\ell}{p} \right\rceil + 1, \ldots, 2\left\lceil \frac{\ell}{p} \right\rceil$ in the second processor and so on. Then the preconditioning step and linked triads do not need any communication between processors. The computation of inner products requires one global communication to accumulate the local inner products computed on each processor. Communication time for computing inner products increases with the number of processors but in general it is small. The matrix-vector multiplication requires exchanging of data between all processors. Communication time for matrix-vector multiplication depends on the size of the matrix and on the number of processors.

4.3.2 Numerical experiments

In this section the results of the numerical experiments are presented. The computations have been executed on a Linux cluster consisting of 4 dual processor PowerPCs with G4 450 MHz processors, 512 MB memory per node. The developed parallel code has been implemented in C and the parallelization has been facilitated using the MPI library, see in [54, 61]. We use the LAPACK library [2] for computing the Cholesky factorization of the preconditioner and for solving the auxiliary linear systems arising in the preconditioned CGM. The optimization options of the compiler have been tuned

to achieve the best performance. Times have been collected using the MPI provided timer. Here the best results from multiple runs are reported.

The first test problem is a class of systems of the form (4.1) with $\ell = 2, 3, \ldots, 10$ equations, where $\mathbf{b}_i = (1, 0)$ and the matrix V is skew-symmetric with elements which are randomly generated constants. Our second test problem comes from the time discretization and Newton linearization of a nonlinear reaction-convection-diffusion system of 10 equations, used in meteorological air-pollution models (cf. [68]). Since the run times here have proved to be very similar to the case of a random 10×10 matrix in the first test problem, we will only present the test results for the first problem.

In what follows, we analyze the obtained parallel time T_p on p processors, relative parallel speed-up $S_p = \frac{T_1}{T_p} \leq p$ and relative efficiency $E_p = \frac{S_p}{p} \leq 1$.

In the experiments we used a stopping criterion $\|r_k\| \leq 10^{-14}$. Table 4.8 shows the required number of iterations.

Tab. 4.8: Number of iterations.

1/h	ℓ									
	1	2	3	4	5	6	7	8	9	10
8	9	10	11	12	12	12	13	13	14	14
16	9	10	12	12	13	13	13	14	14	14
32	9	10	12	12	13	13	14	14	14	14
64	9	10	12	12	13	13	14	14	14	14
128	9	10	12	12	13	13	14	14	14	14

The obtained parallel time T_p on p processors is presented in Tables 4.9 and 4.10. Here ℓ denotes the number of equations. The first column consists of the number of processors. The execution time for problems with $h^{-1} = 32, 64, 128, 192, 256$ in seconds is shown in the next columns. The execution times of the full and truncated version of the algorithm are similar. Because of that we put in Table 4.10 execution times only for systems of 8 and 10 equations. One can see that for relatively small problems, the execution time on one processor is less than one second and parallelization is not necessary. For medium size problems the parallel efficiency on two processors is close to 90% but on three and more processors it decreases. The reason is that communication between two processors in one node is much faster than communication between nodes. For the largest problems ($h^{-1} = 256$) the available physical memory was not enough to solve the problem on one processor. The corresponding numbers in boxes show an atypical progression which is due to the usage of swap memory. The numerical results show that the main advantage of the parallel algorithm is that we can easily solve large problems using a parallel system with distributed memory.

Tab. 4.9: Execution time for full version of GCG-LS.

p	h^{-1}			
	32	64	128	256
$\ell = 2$				
1	0.13	1.06	11.30	130.06
2	0.46	0.99	6.50	69.31
$\ell = 3$				
1	0.22	1.91	19.05	207.86
2	0.55	1.47	13.24	143.40
3	0.60	1.39	8.41	79.30
$\ell = 4$				
1	0.32	2.64	25.62	648.18
2	0.63	1.86	14.43	332.55
3	0.62	1.67	14.58	149.23
4	0.65	1.66	10.05	84.37
$\ell = 5$				
1	0.43	3.44	32.73	912.90
2	0.66	2.26	20.79	216.12
3	0.68	2.10	16.25	153.08
4	0.69	1.95	16.31	155.75
5	0.76	2.06	12.38	94.59
$\ell = 6$				
1	0.54	3.96	39.92	1237.71
2	0.74	2.59	22.10	219.50
3	0.75	2.22	17.15	156.95
4	0.76	2.24	18.09	161.69
5	0.82	2.19	19.06	165.57
6	0.86	2.27	14.98	105.21

p	h^{-1}				
	32	64	128	192	256
$\ell = 7$					
1	0.66	5.13	47.11	171.49	1479.28
2	0.79	3.17	28.60	103.44	667.80
3	0.77	2.74	23.54	82.53	227.45
4	0.82	2.70	19.14	62.73	166.62
5	0.88	3.55	20.95	66.59	361.98
6	0.94	2.80	21.71	68.22	176.53
7	0.97	2.78	18.56	51.21	119.14
$\ell = 8$					
1	0.79	5.96	54.17	306.79	1725.53
2	0.86	3.74	29.99	104.48	771.83
3	0.84	3.30	25.52	86.95	233.69
4	0.86	3.08	19.95	64.44	170.92
5	0.94	3.55	22.14	69.20	178.03
6	1.02	3.62	24.37	73.58	183.49
7	1.07	3.78	25.52	76.36	190.79
8	1.08	4.67	22.30	59.38	132.55
$\ell = 10$					
1	1.08	7.97	70.15	688.04	
2	0.97	4.89	38.64	132.98	1111.04
3	0.95	4.16	32.82	113.15	685.93
4	0.99	4.43	28.75	94.33	248.61
5	1.12	4.13	25.35	76.26	434.87
6	1.18	4.50	27.88	81.52	197.62
7	1.22	4.69	29.99	86.40	205.91
8	1.30	5.49	32.45	92.05	212.42

Tab. 4.10: Execution time for GCG-LS(0).

p	h^{-1}			
	32	64	128	256
$\ell = 8$				
1	0.84	6.07	57.02	2046.74
2	0.48	3.46	31.01	935.01
3	0.51	3.16	26.69	255.81
4	0.59	2.99	21.45	189.93
5	0.67	3.52	23.86	428.05
6	0.76	3.62	26.81	437.50
7	0.82	4.15	29.04	215.17
8	0.85	5.38	26.00	155.73

p	h^{-1}			
	32	64	128	256
$\ell = 10$				
1	1.16	8.51	76.50	
2	0.65	4.87	41.57	1335.88
3	0.67	4.55	36.44	817.74
4	0.71	4.46	32.03	275.20
5	0.86	4.72	29.53	522.18
6	0.96	5.14	32.62	533.91
7	1.06	5.77	35.31	471.83
8	1.09	6.60	38.63	482.45

Figure 4.1 shows the speed-up S_p of the full version of the algorithm obtained for $h^{-1} = 128$ and $\ell = 3, 4, \ldots, 10$. As it was expected when the number of equations ℓ is

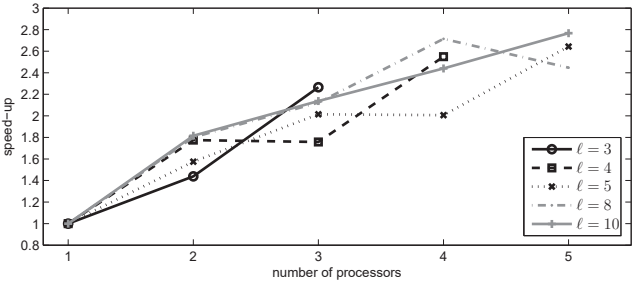

Fig. 4.1: Speed-up of the full version of GCG-LS algorithm.

divisible by the number of processors p the parallel efficiency of the parallel algorithm is higher. The reason is the partitioning of the vectors u_k, d_k, r_k, z_k onto the processors described in previous subsection.

The proposed preconditioner has inherent parallelism—the preconditioning step is implemented without any communications between processors. It has been shown that the code parallelizes well, resulting in a highly efficient treatment of large-scale problems confirmed by the numerical results.

5. OTHER PROBLEMS

In the last chapter we touch upon some related topics where the results of the previous chapters can be used. First we consider the application of nonsymmetric preconditioners to convection-diffusion equations, which can be useful when symmetric operators do not approximate the original operator well, e.g. when the convection term is large (cf. [37]). Then we apply the results of Section 4.2 to nonlinear problems (based on [3, 40]), where the linearized auxiliary equation in the damped inexact Newton method has the form (4.41). Finally a parabolic transport system is considered, where – after time discretization – a nonlinear elliptic system has to be solved on each time level (see [35]).

5.1 Some results on singularly perturbed problems

We consider the iterative solution of large linear systems arising form the discretization of nonsymmetric elliptic problems such as convection-diffusion systems. A preconditioned conjugate gradient method is used, where a nonsymmetric preconditioning operator with constant coefficients is proposed. In this section we study the behaviour of convergence as convection is increasingly dominating. For such convection-dominated problems the suitable choice of preconditioning operator includes nonsymmetric (first order) terms.

Let us consider a general elliptic convection-diffusion BVP

$$\left.\begin{aligned} -\operatorname{div}(A\,\nabla u) + \mathbf{b}\cdot\nabla u + cu &= g \\ u\big|_{\Gamma_D} = 0, \qquad \frac{\partial u}{\partial \nu_A} + \alpha u\big|_{\Gamma_N} &= \gamma, \end{aligned}\right\} \tag{5.1}$$

where Assumptions 3.5 are supposed to be hold, and $g \in L^2(\Omega)$, $\gamma \in L^2(\Gamma_N)$. Then the corresponding operator L has the form (3.25). Further, we introduce the symmetric operator S as defined in (3.26) satisfying Assumptions 3.6 and the energy inner product (3.28). We define the Sobolev space $H^1_D(\Omega) = \{u \in H^1(\Omega) :\ u\big|_{\Gamma_D} = 0\}$ which can be identified with the energy space H_S (see Remark 3.8). Then Assumptions 3.5 ensure that problem (5.1) has a unique weak solution $u \in H^1_D(\Omega)$.

We wish to solve equation (5.1) applying finite element discretization of the problem.

Let $V_h = \operatorname{span}\{\varphi_1, \dots, \varphi_n\} \subset H^1_D(\Omega)$ be a given n dimensional FEM subspace. We seek the FEM solution $u_h \in V_h$, which requires solving the $n \times n$ system

$$\mathbf{L}_h \mathbf{c} = \mathbf{d}_h, \tag{5.2}$$

where $\mathbf{L}_h$ and $\mathbf{d}_h$ are defined in (3.33) and (3.34), respectively. System (5.2) is solved by a proper preconditioned conjugate gradient method. Owing to its nonsymmetry, we use the preconditioned CGN algorithm 2.33. Let us define the nonsymmetric preconditioning operator

$$K \equiv \begin{pmatrix} T \\ V \end{pmatrix}, \quad K \begin{pmatrix} u \\ \eta \end{pmatrix} = \begin{pmatrix} Tu \\ V\eta \end{pmatrix} = \begin{pmatrix} -\operatorname{div}(A\,\nabla u) + \mathbf{w}\cdot\nabla u + zu \\ \dfrac{\partial \eta}{\partial \nu_A} + \zeta\eta\big|_{\Gamma_N} \end{pmatrix} \tag{5.3}$$

for some properly chosen functions $\mathbf{w}, z, \zeta$, where K satisfies Assumptions 3.5 in the obvious sense. Then by Proposition 3.7 the operators $L, K \in BC_S(L^2(\Omega) \times L^2(\Gamma_N))$. Accordingly, the preconditioner for the discretized problem (5.2) is the nonsymmetric stiffness matrix

$$(\mathbf{K}_h)_{ij} = \int_\Omega (A\,\nabla\varphi_i \cdot \nabla\varphi_j + (\mathbf{w}\cdot\nabla\varphi_j)\,\varphi_i + z\varphi_i\varphi_j) + \int_{\Gamma_N} \zeta\varphi_i\varphi_j.$$

Then the preconditioned form of the discrete system (5.2) becomes

$$\mathbf{K}_h^{-1}\mathbf{L}_h\mathbf{c} = \mathbf{K}_h^{-1}\mathbf{d}_h. \tag{5.4}$$

For such preconditioners, it is crucial that systems with $\mathbf{K}_h$ are much cheaper to solve (e.g. with some fast solver) than systems with $\mathbf{L}_h$. This is the case, e.g. if K is symmetric (i.e. $\mathbf{w} = 0$) or if K has constant coefficients. Since the principal parts of L and K coincide, they are compact-equivalent in $H^1_D(\Omega)$ with $\mu = 1$, that is, relation

$$L_S = K_S + Q_S$$

holds in H_S with a compact operator $Q_S \in B(H_S)$, which is defined – similarly to (3.42) – as

$$\begin{aligned} \left\langle Q_S \begin{pmatrix} u \\ u|_{\Gamma_N} \end{pmatrix}, \begin{pmatrix} v \\ v|_{\Gamma_N} \end{pmatrix} \right\rangle_S &= \left\langle L_S \begin{pmatrix} u \\ u|_{\Gamma_N} \end{pmatrix}, \begin{pmatrix} v \\ v|_{\Gamma_N} \end{pmatrix} \right\rangle_S - \left\langle K_S \begin{pmatrix} u \\ u|_{\Gamma_N} \end{pmatrix}, \begin{pmatrix} v \\ v|_{\Gamma_N} \end{pmatrix} \right\rangle_S \\ &= \int_\Omega \big(((\mathbf{b}-\mathbf{w})\cdot\nabla u)v + (c-z)uv\big) + \int_{\Gamma_N} (\alpha-\zeta)uv. \end{aligned} \tag{5.5}$$

Now we apply Algorithm 2.33 for equation (5.4) with $A = \mathbf{K}_h^{-1}\mathbf{L}_h$ and – by calculating the $\mathbf{S}_h$-adjoint of $\mathbf{K}_h^{-1}\mathbf{L}_h$ – with $A^* = \mathbf{S}_h^{-1}\mathbf{L}_h^T\mathbf{K}_h^{-T}\mathbf{S}_h$. Then the following result holds.

Proposition 5.1. *(cf. [10, Thm. 4.3]) Suppose that Assumptions 3.5 hold for the operators L and K (defined in* (3.25) *and* (5.3)*, respectively), and Assumptions 3.6 hold for the operator S (given in* (3.26)*). Let the compact operator Q_S be defined as in* (5.5)*. Let $V_h \subset H_D^1(\Omega)$ be an arbitrary FEM subspace and consider the discrete equation* (5.2) *with the stiffness matrix $\mathbf{K}_h$ as preconditioner. Then the preconditioned CGN algorithm 2.33 converges superlinearly in a mesh independent way, i.e. the residuals satisfy*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \le \frac{2M_K^2}{m_L^2}\left(\frac{1}{k}\sum_{i=1}^{k}\left(\frac{2}{m_K}\sqrt{\lambda_i(Q_S^*Q_S)} + \frac{1}{m_K^2}\lambda_i(Q_S^*Q_S)\right)\right) \xrightarrow{k\to\infty} 0, \quad (5.6)$$

where the positive constants m_L, m_K, M_K come from the S-coercivity and S-boundedness of L and K.

Let us consider the following special case of problem (5.1):

$$\left.\begin{aligned} -\nu\Delta u + \mathbf{b}\cdot\nabla u + cu &= g \\ u\big|_{\Gamma_D} = 0, \qquad \frac{\partial u}{\partial \nu} + \alpha u\Big|_{\Gamma_N} &= \gamma, \end{aligned}\right\} \quad (5.7)$$

where $\nu > 0$ is constant. The coefficient functions $\mathbf{b}$, c, α satisfy Assumptions 3.5. In such problems ν is often small, which means that the problem is convection-dominated. Accordingly, the preconditioning operator (5.3) is

$$K\begin{pmatrix} u \\ u\big|_{\Gamma_N} \end{pmatrix} = \begin{pmatrix} -\nu\Delta u + \mathbf{w}\cdot\nabla u + zu \\ \dfrac{\partial u}{\partial \nu} + \zeta u\Big|_{\Gamma_N} \end{pmatrix}, \quad (5.8)$$

and now we chose $\mathbf{w}, z, \zeta$ to be constant functions. Then systems with $\mathbf{K}_h$ are much cheaper to solve than systems with $\mathbf{L}_h$, e.g. either with multigrid methods or with some fast solver for separable equations on proper domains, see e.g. [56].

The choice of $\mathbf{w}$ is motivated by the following consideration. When ν is small, Theorem 5.1 is not so relevant since it is easy to see that the sequence in (5.6) is proportional to the reciprocal of ν. Although it still tends to zero, this convergence is numerically less relevant since a prescribed accuracy is achieved increasingly later as $\nu \to 0$ (see Tables 3.8 and 3.10 in Chapter 3). The reason is that the above result is based on the symmetric part of K, i.e. it essentially gives the same result if $\mathbf{w} \equiv 0$ or $\mathbf{w}$ is large. Therefore, it is recommended to define $\mathbf{w}$ to be a good constant approximation of $\mathbf{b}$. Then, as $\nu \to 0$, the limit operators of L and K are $\mathbf{b}\cdot\nabla u + cu$ and $\mathbf{w}\cdot\nabla u + zu$,

respectively. To obtain proportional quantities, we assume from now on that $\mathbf{b}$ satisfies and $\mathbf{w}$ is chosen as

$$0 < \beta_1 \le |\mathbf{b}| \le \beta_2, \qquad 0 < \beta_1 \le |\mathbf{w}| \le \beta_2, \tag{5.9}$$

respectively, for some constants β_1, β_2. In fact, if we have coordinatewise $\beta_1^{(i)} := \inf \mathbf{b}_i$ and $\beta_2^{(i)} := \sup \mathbf{b}_i$, then one can define $\mathbf{w}_i := \frac{1}{2}\big(\beta_1^{(i)} + \beta_2^{(i)}\big)$.

For our tests, we consider the following problem:

$$\left.\begin{aligned} -\nu\Delta u + \mathbf{b}\cdot\nabla u + u &= g \\ u\big|_{\partial\Omega} &= 0 \end{aligned}\right\} \tag{5.10}$$

on the unit square $\Omega = [0,1]^2 \subset \mathbb{R}^2$, where $\nu > 0$ and $\mathbf{b} = (\mathbf{b}_1, \mathbf{b}_2)$ is a piecewise constant:

$$\mathbf{b}_1(x,y) := \begin{cases} \lambda & \text{if } 0.5 < y \le 1 \\ 2\lambda & \text{if } 0 \le y \le 0.5, \end{cases} \qquad \mathbf{b}_2(x,y) := \begin{cases} \mu & \text{if } 0 \le x \le 0.5 \\ 2\mu & \text{if } 0.5 < x \le 1. \end{cases}$$

The preconditioning operator is

$$Ku := -\nu\Delta u + \mathbf{w}\cdot\nabla u + u$$

for the same Dirichlet boundary conditions, where the constant vector

$$\mathbf{w} := (1.5\lambda, 1.5\mu)$$

provides an approximation for the first order term of (5.10). To solve (5.10) numerically, we used FEM discretization of the problem with piecewise linear elements and the stopping criterion was

$$q_k := \frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}} \le 10^{-8}$$

for the CGN algorithm 2.33, where $\mathbf{S}_h$ denotes the symmetric part of $\mathbf{L}_h$. The corresponding number of iterations is shown in Table 5.1 for the parameters $\lambda = 1$ and $\mu = 0$ (compare with Tables 3.8 and 3.10).

Although the results in Table 5.1 show that the number of iterations is increasing as ν decreases, it is still reasonable even for small values of ν. Using the symmetric part of L as preconditioner, the convergence remains slow, but much better results can be achieved by using the nonsymmetric preconditioner K. This shows that for singularly perturbed problems the addition of first order terms in the preconditioner improves the performance of the algorithm considerably.

Tab. 5.1: Number of iterations for problem (5.10).

	1/h					
preconditioner:	$\mathbf{S}_h$			$\mathbf{K}_h$		
ν	16	32	64	16	32	64
1	4	4	4	4	4	4
0.1	12	13	13	10	10	10
0.01	53	57	58	18	17	17
0.001	183	239	262	34	31	22
0.0001	308	613	799	50	90	96

5.2 Applications of compact-equivalence to nonlinear problems

The operator pair approach can be applied to nonlinear systems. Here we identify again the spaces $H^1_D(\Omega)^\ell$ and H_S, and the inner product in the product space $H^1_D(\Omega)^\ell$ will be denoted by simply $\langle\cdot,\cdot\rangle_{H^1_D}$.

Consider the nonlinear transport system

$$\left.\begin{aligned} -\operatorname{div}(K_i\,\nabla u_i) + \mathbf{b}_i\cdot\nabla u_i + f_i(x,u_1,\dots,u_\ell) &= g_i \\ u_i\big|_{\Gamma_D} = 0, \qquad K_i\frac{\partial u_i}{\partial \nu} &= \gamma_i \end{aligned}\right\} \qquad (i=1,\dots,\ell) \tag{5.11}$$

on a bounded domain $\Omega\subset\mathbb{R}^d$ ($d=2$ or 3) under the following assumptions:

Assumptions 5.2. *Suppose that*

(i) $\Omega\subset\mathbb{R}^d$ *is a bounded piecewise* C^1 *domain;* Γ_D,Γ_N *are disjoint open measurable subparts of* $\partial\Omega$ *such that* $\partial\Omega=\overline{\Gamma}_D\cup\overline{\Gamma}_N$*;*

(ii) $K_i\in L^\infty(\Omega)$, $\mathbf{b}_i\in C^1(\overline{\Omega})^d$, $g_i\in L^2(\Omega)$ *and* $\gamma_i\in L^2(\Gamma_N)$ $(i=1,\dots,\ell)$*, further, the function* $f=(f_1,\dots,f_\ell):\Omega\times\mathbb{R}^\ell\to\mathbb{R}^\ell$ *is measurable and bounded with respect to the variable* $x\in\Omega$ *and* C^1 *in the variable* $\xi\in\mathbb{R}^\ell$*;*

(iii) there exists $m>0$ *such that* $K_i\ge m$ *holds for all* $i=1,\dots,\ell$*, further,*

$$f'_\xi(x,\xi)\eta\cdot\eta-\frac{1}{2}\left(\max_{1\le i\le\ell}\operatorname{div}\mathbf{b}_i(x)\right)|\eta|^2\ge 0\quad\forall\,(x,\xi)\in\Omega\times\mathbb{R}^d,\ \eta\in\mathbb{R}^d;$$

(iv) let $3\le p$ *(if* $d=2$*) or* $3\le p\le 6$ *(if* $d=3$*), then there exists constants* $c_1,c_2>0$ *such that for any* $(x,\xi_1),(x,\xi_2)\in\Omega\times\mathbb{R}^\ell$

$$\left\|f'_\xi(x,\xi_1)-f'_\xi(x,\xi_2)\right\|\le\left(c_1+c_2\left(\max\{|\xi_1|,|\xi_2|\}\right)^{p-3}\right)|\xi_1-\xi_2|\,.$$

Systems of the form (5.11) arise for instance from the time discretization of nonlinear reaction-convection-diffusion systems. Such systems with homogeneous Dirichlet boundary conditions have been investigated in [3]. The first part of this section is based on that. The proofs given there can be easily modified for the present situation. For brevity, we write (5.11) as

$$\left.\begin{aligned} -\operatorname{div}(\mathbf{K}\,\nabla\mathbf{u}) + \mathbf{b}\cdot\nabla\mathbf{u} + f(x,\mathbf{u}) &= \mathbf{g} \\ \mathbf{u}\big|_{\Gamma_D} = 0, \qquad \mathbf{K}\frac{\partial\mathbf{u}}{\partial\nu} &= \boldsymbol{\gamma} \end{aligned}\right\} \tag{5.12}$$

using vector notations. For any $\mathbf{u} \in H^1_D(\Omega)^\ell$ let

$$\begin{aligned} \langle F(\mathbf{u}), \mathbf{v}\rangle_{H^1_D} &= \int_\Omega \sum_{i=1}^{\ell} \left(K_i\,\nabla u_i\cdot\nabla v_i + (\mathbf{b}_i\cdot\nabla u_i)\,v_i + f_i(x,\mathbf{u})v_i\right) \\ &= \int_\Omega \left(\mathbf{K}\,\nabla\mathbf{u}\cdot\nabla\mathbf{v} + (\mathbf{b}\cdot\nabla\mathbf{u})\cdot\mathbf{v} + f(x,\mathbf{u})\cdot\mathbf{v}\right) \qquad \left(\mathbf{v}\in H^1_D(\Omega)^\ell\right) \end{aligned} \tag{5.13}$$

Owing to Assumptions 5.2 this relation defines a Gâteaux differentiable operator $F : H^1_D(\Omega)^\ell \to H^1_D(\Omega)^\ell$ via the Riesz representation theorem, since for any given $\mathbf{u} \in H^1_D(\Omega)^\ell$ the integral above defines a bounded linear functional on $H^1_D(\Omega)^\ell$. The proof of the theorem below, which relies on the Riesz representation theorem, can be found in [3] for Dirichlet boundary conditions, but it can be easily modified for the present case.

Proposition 5.3. *System* (5.11) *has a unique weak solution, i.e. there exists* $\mathbf{u} \in H^1_D(\Omega)^\ell$ *such that*

$$\langle F(\mathbf{u}), \mathbf{v}\rangle_{H^1_D} = \int_\Omega \mathbf{g}\cdot\mathbf{v} + \int_{\Gamma_N} \boldsymbol{\gamma}\cdot\mathbf{v} \qquad \left(\mathbf{v}\in H^1_D(\Omega)^\ell\right).$$

Let us consider the FEM discretization of (5.13) in the n dimensional FEM subspace $V_h = \operatorname{span}\{\varphi_1,\dots,\varphi_n\} \subset H^1_D(\Omega)$ and we seek the FEM solution $\mathbf{u}_h \in V_h^\ell$:

$$\langle F(\mathbf{u}_h), \mathbf{v}_h\rangle_{H^1_D} = \int_\Omega \mathbf{g}\cdot\mathbf{v}_h + \int_{\Gamma_N} \boldsymbol{\gamma}\cdot\mathbf{v}_h \qquad \left(\mathbf{v}_h\in V_h^\ell\right).$$

The operator $F_h : V_h^\ell \to V_h^\ell$ and the function $\mathbf{f}_h \in V_h^\ell$ are defined by the identities

$$\begin{aligned} \langle F_h(\mathbf{u}_h), \mathbf{v}_h\rangle_{H^1_D} &= \langle F(\mathbf{u}_h), \mathbf{v}_h\rangle_{H^1_D} \qquad \left(\mathbf{v}_h\in V_h^\ell\right), \\ \langle \mathbf{f}_h, \mathbf{v}_h\rangle_{H^1_D} &= \int_\Omega \mathbf{g}\cdot\mathbf{v}_h + \int_{\Gamma_N} \boldsymbol{\gamma}\cdot\mathbf{v}_h \qquad \left(\mathbf{v}_h\in V_h^\ell\right), \end{aligned}$$

thus the problem can be written as a nonlinear algebraic system

$$F_h(\mathbf{u}_h) = \mathbf{f}_h. \tag{5.14}$$

We apply the damped inexact Newton method (DIN) for the iterative solution of problem (5.14). The construction of the DIN method and the related convergence result is as follows.

Let $\mathbf{u}_0 \in V_h^\ell$ be arbitrary. The sequence $(\mathbf{u}_n) \subset V_h^\ell$ is constructed as

Algorithm 5.4 (DIN)**.**

- $\mathbf{u}_{n+1} = \mathbf{u}_n + \tau_n \mathbf{p}_n$, where
- denoting the residual by $\mathbf{r}_h = \mathbf{f}_h - F_h(\mathbf{u}_n)$, the vector $\mathbf{p}_n$ is the solution of

$$\|F_h'(\mathbf{u}_n)\mathbf{p}_n - \mathbf{r_h}\|_{H_D^1} \le \delta_n \|\mathbf{r}_h\|_{H_D^1} \quad \text{with } 0 < \delta_n \le \delta_0 < 1,$$

- $\tau_n = \min\left\{1, \dfrac{1-\delta_n}{(1+\delta_n)^2} \dfrac{m^2}{L\,\|F_h(\mathbf{u}_n) - \mathbf{f}_h\|_{H_D^1}}\right\}$.

Theorem 5.5. *Let Assumptions 5.2 hold. If* $\delta_n \le \text{const} \cdot \|F_h(\mathbf{u}_n) - \mathbf{f}_h\|_{H_D^1}^\gamma$ *with some* $0 < \gamma \le 1$*, then the convergence is locally of order* $1+\gamma$*, that is the convergence is linear for* n_0 *steps until* $\|F_h(\mathbf{u}_n) - \mathbf{f}_h\|_{H_D^1}^\gamma \le \varepsilon$*, where* $\varepsilon \le (1-\delta_0)\frac{m^2}{2L}$ *(here and in the definition of* τ_n *the constant* L *comes from the Lipschitz continuity of* F'*), and further on (as* $\tau_n \equiv 1$*)*

$$\|\mathbf{u}_n - \mathbf{u}_h\|_{H_D^1} \le d_1 q^{(1+\gamma)^{n-n_0}}$$

with some $d_1 > 0$*,* $0 < q < 1$*, which provides mesh independent convergence rate for the DIN method.*

It can be shown that the conditions of [22, Thm. 5.12] are satisfied. This has been done for Dirichlet boundary conditions in [3] and that argument can also be applied to the present case with minor modifications. In each step the construction of $\mathbf{u}_n$ requires the approximate solution of the linearized problem

$$F_h'(\mathbf{u}_n)\mathbf{p}_h = \mathbf{r}_n, \tag{5.15}$$

which is equivalent to the FEM solution in V_h^ℓ of the linear elliptic system

$$\left.\begin{aligned} -\operatorname{div}(K_i\,\nabla p_i) + \mathbf{b}_i \cdot \nabla p_i + \sum_{j=1}^{\ell} \partial_j f_i(x, \mathbf{u}_n) p_j &= r_i \\ p_i\big|_{\Gamma_D} = 0, \qquad K_i \frac{\partial p_i}{\partial \nu} &= \varrho_i \end{aligned}\right\} \qquad (i = 1, \dots, \ell) \tag{5.16}$$

where

$$r_i = g_i + \operatorname{div}(K_i \nabla u_{n,i}) - \mathbf{b}_i \cdot \nabla u_{n,i} - f_i(x, \mathbf{u}_n) \quad \text{and} \quad \varrho_i = \gamma_i - K_i \frac{\partial u_{n,i}}{\partial \nu}.$$

Denoting by $\mathbf{c}$ and $\mathbf{d}$ the coefficient vectors of $\mathbf{p}_h$ and $\mathbf{r}_h$, and by $\mathbf{L}_h^{(n)}$ the stiffness matrix corresponding to (5.16), equation (5.15) requires the solution of the linear algebraic system

$$\mathbf{L}_h^{(n)} \mathbf{c} = \mathbf{d}. \tag{5.17}$$

The equivalent operator framework of Section 4.2 can be applied to the auxiliary linear problem (5.16), since it has the form (4.38). The preconditioner for the discrete system (5.17) is defined as the stiffness matrix $\mathbf{S}_h$ of S in $H_D^1(\Omega)^\ell$, where S is defined as in (4.44)-(4.45) with $G_i = K_i$. Then we apply the CGN algorithm 2.33 for the preconditioned system

$$\mathbf{S}_h^{-1} \mathbf{L}_h^{(n)} \mathbf{c} = \mathbf{S}_h^{-1} \mathbf{d}.$$

Combining the convergence results for the CGN and the DIN algorithms 2.33 and 5.4, the combined iteration provides mesh independent convergence, with superlinear convergence rate for both the inner and outer iterations (see Corollary 4.12 and Theorem 5.5). Moreover, the operators S_i are decoupled, hence in each Newton step the linearized system (5.16) is preconditioned by an ℓ-tuple of independent symmetric elliptic operators.

5.3 *A convergent time discretization scheme for nonlinear parabolic transport systems*

Nonlinear parabolic systems arise in various mathematical models where transport type processes are involved, and their numerical solution is a challenging task ([68]). This is both due to the compound nature of the equations that involve second, first and zeroth-order terms (i.e. describing diffusion, convection and reaction type parts of the process), and the large size of the problem that comes both from the possibly huge number of equations and from the discretization.

In this section we introduce an approach combining time discretization with outer-inner iterations, proposed for the finite element discretization of the problem. The outer-inner iterations for the elliptic subproblems involve the damped inexact Newton and the preconditioned conjugate gradient methods (PCG), exploiting their superlinear convergence properties, based on [3, 10]. First we describe the problem, then some numerical experiments are presented for reaction-convection-diffusion systems from air

pollution models.

We consider systems of the form

$$\left.\begin{aligned} \frac{\partial u_i}{\partial t} - \operatorname{div}(K_i(x)\,\nabla u_i) + \mathbf{b}_i \cdot \nabla u_i + c_i(x) u_i + f_i(x,t,u_1,\dots,u_\ell) &= 0 \\ u_i(x,0) = \varphi_i(x) \quad (x \in \Omega), \qquad u_i\big|_{\partial\Omega\times\mathbb{R}^+} &= 0, \end{aligned}\right\} \tag{5.18}$$

$(i = 1, \dots, \ell)$, under the following assumptions:

Assumptions 5.6. *Suppose that*

(i) the bounded domain $\Omega \subset \mathbb{R}^d$ is C^2-diffeomorphic to a convex domain;

(ii) for all $i = 1, \dots, \ell$ the functions $K_i \in C^1(\overline{\Omega})$ and $\mathbf{b}_i \in C^1(\overline{\Omega})^d$, further, the function $f = (f_1, \dots, f_\ell) : \Omega \times \mathbb{R}^+ \times \mathbb{R}^\ell \to \mathbb{R}^\ell$ is measurable and bounded with respect to the variable $x \in \Omega$ and C^1 in the variables $t \geq 0$ and $\xi \in \mathbb{R}^\ell$;

(iii) there exists $m > 0$ such that

$$K_i \geq m \quad \text{and} \quad c_i - \frac{1}{2}\operatorname{div} \mathbf{b}_i \geq 0 \qquad (i = 1, \dots, \ell);$$

(iv) there exists $c_0 > 0$ such that

$$f'_\xi(x,\xi)\eta \cdot \eta - \frac{1}{2}\left(\max_{1\leq i\leq \ell} \operatorname{div} \mathbf{b}_i(x)\right)|\eta|^2 \geq -c_0\,|\eta|^2 \quad \forall\ (x,\xi) \in \Omega \times \mathbb{R}^d,\ \eta \in \mathbb{R}^d;$$

(v) let $p^* := +\infty$ (if $d = 2$) or $p^* := \frac{2d}{d-2}$ (if $d > 2$, where d is the space dimension). Then there exist constants $c_1 \geq 0$ and $\alpha \leq \frac{p^*}{d}$ such that for any $x \in \Omega$, $\xi_1, \xi_2 \in \mathbb{R}^\ell$ and $t \geq 0$

$$|f(x,t,\xi_1) - f(x,t,\xi_2)| \leq c_1\,(1 + \max\{|\xi_1|^\alpha, |\xi_2|^\alpha\})\,|\xi_1 - \xi_2|\,;$$

(vi) $\varphi_i \in C(\overline{\Omega})$ for all $i = 1, \dots, \ell$.

Systems of the form (5.18) arise e.g. in nonlinear reaction-convection-diffusion systems such as air pollution models [68], where f_i describe the rate of chemical reactions. Here typically

$$f_i(x,t,u_1,\dots,u_\ell) = \sum_{j=1}^{\ell} c_{ij} u_i u_j, \tag{5.19}$$

in which case $\alpha = 1$ in assumption (iv).

For brevity, using obvious vector notations, (5.18) can be written as

$$\left.\begin{aligned} \frac{\partial \mathbf{u}}{\partial t} + L\mathbf{u} + f(x,t,\mathbf{u}) &= \mathbf{0} \\ \mathbf{u}(x,0) &= \varphi(x), \end{aligned}\right\} \tag{5.20}$$

where

$$L\mathbf{u} := -\operatorname{div}(\mathbf{K}\,\nabla\mathbf{u}) + \mathbf{b}\cdot\nabla\mathbf{u} + \mathbf{c}\mathbf{u} \qquad \text{for } \mathbf{u} \in \left(H^2(\Omega)\cap H_0^1(\Omega)\right)^\ell. \tag{5.21}$$

Now some numerical results are presented. Let $\Omega \subset \mathbb{R}^2$ be the unit square and $K_i \equiv 1$ $(i = 1,\dots,\ell)$ in (5.18), i.e. for simplicity only the case of Laplacian is considered as the principal part of the elliptic operators. Having chosen the convection term to be $\mathbf{b} = (1,1)$, the following type of equations are used for the numerical tests:

$$\left.\begin{aligned} &\frac{\partial u_i}{\partial t} - \Delta u_i + \frac{\partial u_i}{\partial x} + \frac{\partial u_i}{\partial y} + f_i(x,y,t,u_1,\dots,u_\ell) = 0 \\ &u_i(x,y,0) = \varphi_i(x) \quad ((x,y)\in[0,1]^2) \\ &u_i\big|_{\partial\Omega\times\mathbb{R}^+} = 0 \end{aligned}\right\} \qquad (i = 1,\dots,\ell), \tag{5.22}$$

where a bounded time interval $[0,T]$ is considered with maximal time $T = 1$. The initial function φ is a polynomial satisfying the boundary conditions. The nonlinear terms in (5.22) have the form

$$f(x,y,t,\mathbf{u}) = 4\mathbf{A}\,|\mathbf{u}|^2\,\mathbf{u},$$

where $\mathbf{A} \in \mathbb{R}^{\ell\times\ell}$ is a lower triangular matrix with all 1 entries. This specific choice ensures that $f'_\xi(x,y,t,\xi)$ is positive definite. In the first experiment an additional term has to be added, since an exact solution has to be known to be able to compute the errors.

In the following tables the number of outer DIN iterations executed is every time step and the number of outer PCG iterations carried out in each DIN step are denoted by n and n_{inn}, respectively. The stopping criterion in the DIN method was chosen to be $\|F_h(u) - b_h\| < 10^{-8}$.

First the results of an experiment with 4 equations are presented, with the emphasis of the mesh independent convergence of the numerical solutions. The exact solutions of (5.18) were chosen in the form

$$u^*(x,y) = C\cdot\left(x - x^2\right)\left(y - y^2\right)e^{-t}, \tag{5.23}$$

thus another term was added to the nonlinear term f_i.

In Table 5.2 the errors are shown in four different points in the time interval, when various spatial ($h = 1/N$) and time parameters (τ) were chosen.

Tab. 5.2: First order convergence in τ for 4 equations.

		error $= \|u_h - u^*\|$			
t	τ	$N = 8$	$N = 16$	$N = 32$	$N = 64$
0.25	1/4	0.01111	0.01111	0.01111	0.01111
	1/8	0.00597	0.00595	0.00595	0.00595
	1/16	0.00310	0.00307	0.00306	0.00306
	1/32	0.00159	0.00156	0.00155	0.00154
0.50	1/4	0.01063	0.01061	0.01060	0.01060
	1/8	0.00518	0.00515	0.00515	0.00514
	1/16	0.00254	0.00252	0.00251	0.00251
	1/32	0.00127	0.00125	0.00124	0.00124
0.75	1/4	0.00862	0.00859	0.00858	0.00859
	1/8	0.00408	0.00405	0.00405	0.00405
	1/16	0.00199	0.00197	0.00196	0.00195
	1/32	0.00099	0.00097	0.00097	0.00096
1.00	1/4	0.00677	0.00675	0.00674	0.00674
	1/8	0.00318	0.00316	0.00316	0.00316
	1/16	0.00155	0.00153	0.00153	0.00152
	1/32	0.00077	0.00076	0.00075	0.00075

Considering the rows, it can be seen that the error is independent of the choice of the spatial parameter, thus the convergence is mesh independent. Picking up one of the time levels t in the interval $[0, T]$ from the first column of Table 5.2, it is obvious that halving the time parameter τ causes the halving of the errors, thus $\mathcal{O}(\tau)$ accuracy can be obtained in this procedure with respect to time.

Tab. 5.3: Number of DIN and inner PCG steps for 4 equations, tolerance level $= 10^{-8}$.

$N = h^{-1} = 32$											
$t = 0.00$			$t = 0.25$			$t = 0.50$			$t = 0.75$		
n	$\|r_h\|_{\mathbf{S}_h}$	n_{inn}	n	$\|r_h\|_{\mathbf{S}_h}$	n_{inn}	n	$\|r_h\|_{\mathbf{S}_h}$	n_{inn}	n	$\|r_h\|_{\mathbf{S}_h}$	n_{inn}
0	0.34450858	1	0	0.27899503	1	0	0.21905425	1	0	0.17089308	1
1	0.10386448	2	1	0.08408514	2	1	0.06601429	2	1	0.05149946	2
2	0.00915977	2	2	0.00743173	2	2	0.00584812	2	2	0.00457164	2
3	0.00007365	4	3	0.00004867	5	3	0.00003060	5	3	0.00001933	5
4	0.00000045	5	4	0.00000024	5	4	0.00000012	5	4	0.00000006	4
5	0.00000000	-	5	0.00000000	-	5	0.00000000	-	5	0.00000000	-

In every time step consecutive DIN iterations have to be carried out until an acceptable residual error is reached, where in each step an auxiliary equation has to be

solved using a PCG algorithm for the normalized equation. Thus in the nth DIN step the residual error $\|r_h\|_{\mathbf{S}_h}$ was checked first, then a PCG was carried out n_{inn} times. The results for four equations can be seen in Table 5.3.

Tab. 5.4: Number of DIN and inner PCG steps for 10 equations, tolerance level = 10^{-8}.

$N = h^{-1} = 32$											
$t = 0.00$			$t = 0.25$			$t = 0.50$			$t = 0.75$		
n	$\|r_h\|_{\mathbf{S}_h}$	n_{inn}	n	$\|r_h\|_{\mathbf{S}_h}$	n_{inn}	n	$\|r_h\|_{\mathbf{S}_h}$	n_{inn}	n	$\|r_h\|_{\mathbf{S}_h}$	n_{inn}
0	0.09482921	2	0	0.01575960	2	0	0.00260900	3	0	0.00043107	4
1	0.02841575	2	1	0.00472253	3	1	0.00078171	4	1	0.00012916	4
2	0.00254771	3	2	0.00042270	4	2	0.00006996	5	2	0.00001156	5
3	0.00000222	7	3	0.00000024	6	3	0.00000003	5	3	0.00000000	-
4	0.00000000	-	4	0.00000000	-	4	0.00000000	-	-	-	-

Table 5.4 shows the results for a system of convection-diffusion consisting of 10 equations, where the system is derived from an air pollution model (cf. [68]), from which the linearized system (4.37) is originated. The residual errors follow the same pattern as for the smaller problem. Since no exact solution is available, only the approximate solutions calculated in a pair of grids can be compared, when τ and $\tau/2$ are used as time parameters. The results are shown in Table 5.5 which exhibit the numerical convergence of the algorithm.

Tab. 5.5: Error estimation in τ for 10 equations.

		$\left\\| u_h^{(\tau)} - u_h^{(\tau/2)} \right\\|$			
t	τ	$N = 8$	$N = 16$	$N = 32$	$N = 64$
0.25	1/4	5.6032e-03	5.5971e-03	5.5962e-03	5.6078e-03
	1/8	2.9354e-03	2.9157e-03	2.9357e-03	2.9311e-03
	1/16	1.3272e-03	1.3174e-03	1.3210e-03	1.3189e-03
0.50	1/4	1.5072e-03	1.4957e-03	1.4987e-03	1.4979e-03
	1/8	3.9029e-04	3.8588e-04	3.8338e-04	3.8192e-04
	1/16	8.9336e-05	8.7142e-05	8.6723e-05	8.6821e-04
0.75	1/4	3.0803e-04	3.0438e-04	3.0280e-04	3.0129e-04
	1/8	3.9768e-05	3.8658e-05	3.8191e-05	3.7851e-05
	1/16	4.5972e-06	4.3512e-06	4.2254e-06	4.1974e-06
1.00	1/4	5.7434e-05	5.6288e-05	5.5750e-05	5.5580e-05
	1/8	3.7062e-06	3.5447e-06	3.4740e-06	3.4536e-06
	1/16	2.1499e-07	1.9754e-07	1.9221e-07	1.8993e-07

SUMMARY

The numerical solution of linear elliptic partial differential equations consists of two main steps: discretization and iteration, where generally some conjugate gradient method is used for solving the finite element discretization of the problem. However, when for elliptic problems the discretization parameter tends to zero, the required number of iterations for a prescribed tolerance tends to infinity. The remedy is suitable preconditioning, which can rely on Hilbert space theory. The subject of this thesis is the investigation and numerical realization of the existing theory of operator preconditioning, and the extension of the theoretical results to cases that have not been covered before. Operator preconditioning means that the preconditioning process takes place on the operator level, that is, we look for a suitable preconditioning operator for the operator equation – based on the theory of equivalent operators – and then we use its discretization as a preconditioner for the discrete system. In this thesis we have primarily dealt with symmetric preconditioners. The main results are the following.

In Chapter 3 we have first investigated the theoretical results for convection-diffusion equations with homogeneous mixed boundary conditions. We have shown that the numerical computations provide better results than the theoretical estimate. The convergence rate has remained valid even in cases that are not covered by the theory. Then we have extended the theory to the nonhomogeneous case using operator pairs and we have obtained an analogous mesh independent convergence result as in the homogeneous case. We have derived a similar convergence estimate in the finite difference case for a special model problem.

In Chapter 4 we have extended the mesh independent superlinear convergence results from equations to systems. An important advantage of the proposed preconditioning method for systems is that one can define decoupled preconditioners, thus parallelization of the auxiliary systems is available. We have developed and implemented an efficient parallel algorithm for decoupled symmetric preconditioners.

In Chapter 5 we have discussed some related problems where the considered preconditioning approach can be used. We have shown that the use of nonsymmetric preconditioners is more advantageous for singularly perturbed problems than symmetric preconditioning. The application of the results of the preceding chapters to nonlinear elliptic and parabolic problems closes the dissertation.

MAGYAR NYELVŰ ÖSSZEFOGLALÁS

Lineáris elliptikus parciális differenciálegyenletek numerikus megoldásának két fő lépése a diszkretizáció és iteráció. Az esetek nagyrészében egy végeselem-módszerrel kapott nagyméretű lineáris algebrai egyenletrendszert oldunk meg iterációs eljárással, például valamilyen konjugált gradiens-módszerrel. A rácsfelosztás finomításával azonban egy adott pontossághoz szükséges iterációk száma végtelenhez tart. A probléma megoldása a prekondicionálásnak nevezett eljárás, amely Hilbert-terek operátorainak elméletére is támaszkodik. A dolgozat tárgya az operátor-prekondicionálás néhány ismert eredményének vizsgálata és numerikus megvalósítása, továbbá az elmélet kiterjesztése eddig még nem tárgyalt esetekre. Itt a prekondicionálás operátorszinten történik, vagyis az adott elliptikus operátoregyenlethez keresünk egy másik alkalmas elliptikus operátort, amelynek a diszkretizáltját használjuk prekondicionáló mátrixként a diszkrét egyenlethez. Ebben a dolgozatban elsősorban szimmetrikus prekondicionáló operátorokkal foglalkoztunk. A témában elért fő eredmények a következők.

A 3. fejezetben elliptikus konvekció-diffúzió egyenleteket vizsgáltunk homogén harmadfajú peremfeltétel mellett. Megmutattuk, hogy a numerikus számítások jobb eredményeket adnak, mint az elméleti becslések, sőt, a konvergencia gyorsasága az elmélet által le nem fedett esetekben is érvényben maradt. Ezt követően operátor-párok alkalmazásával kiterjesztettük az elméletet az inhomogén peremfeltétel esetére. Végül hasonló konvergenciabecslést bizonyítottunk véges differenciás diszkretizáció esetén egy speciális modellfeladatra.

A 4. fejezetben kiterjesztettük az egyenletekre elért rácsfüggetlen szuperlineáris konvergenciaeredményeket rendszerekre. A vizsgált prekondicionáló eljárás különösen előnyös tulajdonsága, hogy széteső szimmetrikus prekondicionáló operátor használata esetén a keletkező segédfeladatok kezelése egymástól független egyenletek megoldását jelenti, amely jól párhuzamosítható. A fejezet végén bemutattunk és alkalmaztunk egy ilyen prekondicionáló operátortípusra kifejlesztett párhuzamos algoritmust.

Az 5. fejezetben néhány olyan problémát érintettünk röviden, ahol az eddig tárgyalt eljárások felhasználhatók. Megmutattuk, hogy szinguláris perturbált feladatokra a nemszimmetrikus prekondicionáló operátorok jóval hatékonyabbak, mint a szimmetrikusak. Zárásként az előző fejezetek eredményeit alkalmaztuk nemlineáris elliptikus, illetve parabolikus feladatokra.

BIBLIOGRAPHY

[1] R. A. Adams. *Sobolev Spaces*. Academic Press, 1975.

[2] E. Anderson et al. *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics, third edition, 1999.

[3] I. Antal and J. Karátson. A mesh independent superlinear algorithm for some nonlinear nonsymmetric elliptic systems. *Comput. Math. Appl.*, 55:2185–2196, 2008.

[4] S. F. Ashby, T. A. Manteuffel, and P. E. Saylor. A taxonomy for conjugate gradient methods. *SIAM J. Numer. Anal.*, 27:1542–1568, 1990.

[5] K. Atkinson and W. Han. *Theoretical Numerical Analysis: a Functional Analysis Framework*. Number 39 in Texts in Applied Mathematics. Springer, second edition, 2005.

[6] O. Axelsson. A generalized conjugate gradient least square method. *Numer. Math.*, 51:209–227, 1987.

[7] O. Axelsson. *Iterative Solution Methods*. Cambridge University Press, 1994.

[8] O. Axelsson and J. Karátson. Symmetric part preconditioning for the conjugate gradient method in Hilbert space. *Numer. Funct. Anal.*, 24 (No. 5-6):455–474, 2003.

[9] O. Axelsson and J. Karátson. Superlinearly convergent CG methods via equivalent preconditioning for nonsymmetric elliptic operators. *Numer. Math.*, 99:197–223, 2004.

[10] O. Axelsson and J. Karátson. Mesh independent superlinear PCG rates via compact-equivalent operators. *SIAM J. Numer. Anal.*, 45 (4):1495–1516, 2007.

[11] O. Axelsson and J. Karátson. Equivalent operator preconditioning for elliptic problems. *Numer. Algor.*, 50:297–380, 2009.

[12] M. Benzi. Preconditioning techniques for large linear systems: a survey. *J. of Comp. Phys.*, 182:418–477, 2002.

[13] P. G. Ciarlet. *The Finite Element Method for Elliptic Problems*. North-Holland, 1978.

[14] P. Concus and G. H. Golub. A generalized conjugate gradient method for nonsymmetric systems of linear equations. In R. Glowinski and J. Lions, editors, *Computing Methods in Applied Sciences and Engineering*, volume 134 of *Lecture Notes in Economics and Math. Syst.*, pages 56–65. Springer, 1976.

[15] J. B. Conway. *A Course in Functional Analysis*. Number 96 in Graduate Texts in Mathematics. Springer, 1990.

[16] L. Czách. The steepest descent method for elliptic differential equations (in Russian). C.Sc. thesis, 1955.

[17] J. W. Daniel. The conjugate gradient method for linear and nonlinear operator equations. *SIAM J. Numer. Anal.*, 4:10–26, 1967.

[18] E. G. D'yakonov. On an iterative method for the solution of finite difference equations (in Russian). *Dokl. Akad. Nauk. SSSR*, 138:522–525, 1961.

[19] S. C. Eisenstat, H. C. Elman, and M. H. Schultz. Variational iterative methods for nonsymmetric systems of linear equations. *SIAM J. Numer. Anal.*, 20 (No. 2):345–357, 1983.

[20] V. Faber and T. A. Manteuffel. Necessary and sufficient conditions for the existence of a conjugate gradient method. *SIAM J. Numer. Anal.*, 21:352–362, 1984.

[21] V. Faber, T. A. Manteuffel, and S. V. Parter. On the theory of equivalent operators and application to the numerical solution of uniformly elliptic partial differential equations. *Advances in Applied Mathematics*, 11:109–163, 1990.

[22] I. Faragó and J. Karátson. *Numerical Solution of Nonlinear Elliptic Problems via Preconditioning Operators: Theory and Applications*, volume 11 of *Advances in Computation: theory and practice*. NOVA Science Publishers, 2002.

[23] C. I. Goldstein, T. A. Manteuffel, and S. V. Parter. Preconditioning and boundary conditions without H_2 estimates: L_2 condition numbers and the distribution of the singular values. *SIAM J. Numer. Anal.*, 30 (No. 2):343–376, 1993.

[24] J. E. Gunn. The numerical solution of $\nabla \cdot a \nabla u = f$ by a semi-explicit alternating direction iterative method. *Numer. Math.*, 6:181–184, 1964.

[25] W. Hackbush. *Multigrid Methods and Applications*, volume 4 of *Springer Series in Computational Mathematics*. Springer, 1985.

[26] W. Hackbush. *Elliptic Differential Equations: Theory and Numerical Treatment*, volume 18 of *Springer Series in Computational Mathematics*. Springer, 1992.

[27] R. M. Hayes. Iterative methods of solving linear problems in Hilbert space. *Nat. Bur. Standards Appl. Math. Ser.*, 39:71–104, 1954.

[28] M. R. Hestenes and E. Stiefel. Methods of conjugate gradient for solving linear systems. *J. Res. Natl. Bur. Stand.*, 49:409–436, 1952.

[29] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1986.

[30] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991.

[31] W. Joubert, T. A. Manteuffel, S. V. Parter, and S. Wong. Preconditioning second-order elliptic operators: Experiment and theory. *SIAM J. Sci. Stat. Comput.*, 13:259–288, 1992.

[32] J. Kadlec. On the regularity of the solution of the Poisson problem on a domain with boundary locally similar to the boundary of a convex open set. *Czechosl. Math. J.*, 14 (89):386–393, 1964.

[33] L. V. Kantorovich and G. Akilov. *Functional Analysis in Normed Spaces*. Pergamon Press, second edition, 1982.

[34] J. Karátson. Superlinear PCG algorithms: symmetric part preconditioning and boundary conditions. *Numer. Funct. Anal.*, 29:590–611, 2008.

[35] J. Karátson and T. Kurics. A convergent time discretization scheme for nonlinear parabolic transport systems. `http://www.cs.elte.hu/applanal/preprints/parextd.ps`, 2007.

[36] J. Karátson and T. Kurics. Superlinearly convergent PCG algorithms for some nonsymmetric elliptic systems. *J. Comp. Appl. Math.*, 212 (2):214–230, 2008.

[37] J. Karátson and T. Kurics. Some superlinear PCG methods for discretized elliptic problems. In G. Maroulis and T. E. Simos, editors, *American Institute of Physics Conference Series*, volume 1148, pages 861–864, 2009.

[38] J. Karátson and T. Kurics. Superlinear PCG methods for FDM discretizations of convection-diffusion equations. In S. Margenov, L. G. Vulkov, and J. Waśniewski, editors, *Numerical Analysis and Its Applications*, volume 5434 of *LNCS*, pages 345–352. Springer, 2009.

[39] J. Karátson, T. Kurics, and I. Lirkov. A parallel algorithm for systems of convection-diffusion equations. In T. Boyanov et al., editors, *Numerical Methods and Applications*, volume 4310 of *LNCS*, pages 65–73. Springer, 2007.

[40] T. Kurics. Equivalent operator preconditioning for elliptic problems with nonhomogeneous mixed boundary conditions. To appear in Computers and Mathematics with Applications.

[41] T. Kurics. On symmetric part PCG for mixed elliptic problems. In I. Lirkov, S. Margenov, and J. Waśniewski, editors, *Large Scale Scientific Computing*, volume 3743 of *LNCS*, pages 679–686. Springer, 2006.

[42] T. Kurics. On the superlinear convergence of PCG algorithms: numerical experiments for convection-diffusion equations. *Computers and Mathematics with Applications*, 55/10:2318–2328, 2008.

[43] C. Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Res. Natl. Bur. Stand.*, 45:255–282, 1950.

[44] C. Lanczos. Solution of systems of linear equations by minimized iterations. *J. Res. Natl. Bur. Stand.*, 49:33–53, 1952.

[45] T. A. Manteuffel and J. Otto. Optimal equivalent preconditioners. *SIAM J. Numer. Anal.*, 30 (No. 3):790–812, 1993.

[46] T. A. Manteuffel and S. V. Parter. Preconditioning and boundary conditions. *SIAM J. Numer. Anal.*, 27 (No. 3):656–694, 1990.

[47] W. M. Patterson. *Iterative Methods for the Solution of a Linear Operator Equation in Hilbert Space – A Survey.* Number 394 in Lecture Notes in Mathematics. Springer, 1974.

[48] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical Mathematics.* Springer, 2000.

[49] A. Quarteroni and A. Valli. *Numerical Approximation of Partial Differential Equations.* Number 23 in Computational Mathematics. Springer, 1997.

[50] F. Riesz and B. Sz.-Nagy. *Functional Analysis.* Dover, 1990.

[51] T. Rossi and J. Toivanen. A parallel fast direct solver for block tridiagonal systems with separable matrices of arbitrary dimension. *SIAM J. Sci. Comput.*, 20 (No. 5):1778–1796, 1999.

[52] Y. Saad. *Iterative Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics, 2003.

[53] Y. Saad and M. H. Schultz. Conjugate gradient-like algorithms for solving nonsymmetric linear systems. *Math. Comput.*, 44:417–424, 1985.

[54] M. Snir, S. Otto, S. Huss-Lederman, D. Walker, and J. Dongara. *MPI: The complete reference*. Scientific and engineering computation. The MIT Press, 1997. Second printing.

[55] J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis*. Number 12 in Texts in Applied Mathematics. Springer, second edition, 1993.

[56] P. N. Swarztrauber. A direct method for the discrete solution of separable elliptic equations. *SIAM J. Numer. Anal.*, 11:1136–1150, 1974.

[57] P. N. Swarztrauber. The methods of cyclic reduction, Fourier analysis and the FACR algorithm for the discrete solution of Poisson's equation on a rectangle. *SIAM Rev.*, 19 (No. 3):490–501, 1977.

[58] H. A. van der Vorst. Iterative solution methods for certain sparse linear systems with a nonsymmetric matrix arising from PDE-problems. *J. Comput. Phys.*, 44:1–19, 1981.

[59] H. A. van der Vorst. *Iterative Krylov Methods for Large Linear Systems*. Cambridge University Press, 2003.

[60] R. S. Varga. *Matrix Iterative Analysis*. Springer, second edition, 2000.

[61] D. Walker and J. Dongara. MPI: a standard Message Passing Interface. *Supercomputer*, 63:56–68, 1996.

[62] O. Widlund. A Lanczos method for a class of non-symmetric systems of linear equations. *SIAM J. Numer. Anal.*, 15:801–812, 1978.

[63] R. Winther. Some superlinear convergence results for the conjugate gradient method. *SIAM J. Numer. Anal.*, 17:14–17, 1980.

[64] K. Yosida. *Functional Analysis*. Springer, sixth edition, 1980.

[65] D. M. Young. *Iterative Solution of Large Linear Systems*. Academic Press, 1971.

[66] E. Zeidler. *Nonlinear Functional Analysis and its Applications. II/A: Linear Monotone Operators*. Springer, 1990.

[67] E. Zeidler. *Nonlinear Functional Analysis and its Applications. II/B: Nonlinear Monotone Operators*. Springer, 1990.

[68] Z. Zlatev. *Computer Treatment of Large Air Pollution Models*. Kluwer Academic Publishers, 1995.